DOCUMENT RESUME

ED 200 219 IR 009 253

AUTHOR	Vondran, Raymond F.
TITLE	National Union Catalog Experience: Implications for Network Planning. Network Planning Paper No. 6.
INSTITUTION	Library of Congress, Washington, D.C. Network Development Office.
SPONS AGENCY	National Commission on Libraries and Information Science, Washington, D. C.
REPORT NO	ISBN-0-8444-0325-3
PUB DATE	80
NOTE	64p.
AVAILABLE FROM	Customer Services Section, Cataloging Distribution Service, Library of Congress, Navy Yard Annex, Bldg. 159, Washington, DC 20541.
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Cataloging; Flow Charts; Library Networks; *Operations Research; *Publications; Records (Forms); Subject Index Terms; Tables (Data); *Union Catalogs
IDENTIFIERS	Library of Congress; *National Union Catalog

ABSTRACT

This study conducted to provide background data for the systematic development of the library bibliographic component of a national network analyzes the procedures used in producing the National Union Catalog (NUC), the nationwide union catalog in card form maintained at the Library of Congress (LC), and examines the variations found in records as they are integrated into the file. The report discusses NUC operations and workflow, characteristics of reporting libraries, differences in headings between outside reports to the NUC and LC records, and differences in bibliographic description between outside reports to the NUC and LC records. Conclusions for a nationwide database derived from the analysis of processes in the catalog publication division are presented along with recommendations for future investigation. Tables and figures are provided. (FM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED200219

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ISSN 0160-9742

NETWORK PLANNING PAPER

NETWORK DEVELOPMENT OFFICE

NUMBER 6

DATE 1980

TITLE National Union Catalog Experience: Implications for Network Planning

AUTHOR Raymond F. Vondran

Funded by the National Commission on Libraries and Information Science

Library of Congress

Washington

IR009253

Library of Congress Cataloging in Publication Data

Vondran, Raymond F 1946-
National Union Catalog experience.

(Network planning paper ; no. 6 ISSN 0160-9742)
Includes bibliographical references.
Supt. of Docs. no.: LC 1.37:6
1. United States. Library of Congress. National union catalog. I. United States. Library of Congress. Network Development Office. II. Title. III. Series.
Z881.A1U687 025.3'13 79-607146
ISBN 0-8444-0325-3

Available by request from the Customer Services Section, Cataloging Distribution Service, Library of Congress, Navy Yard Annex, Bldg. 159, Washington, D.C. 20541

FOREWORD

This report presents a description of procedures used in producing the National Union Catalog, a nationwide union catalog in card form that has been maintained at the Library of Congress since 1901, and an analysis of the variations found in records as they are integrated into the file. In order to understand some of the problems involved and derive some of the requirements for producing a consistent nationwide library data base, it appeared useful to study the procedures employed in building one of the largest files with the attributes of a catalog that is maintained today.

This study was funded by the National Commission on Libraries and Information Science and performed under contract by Dr. Raymond F. Vondran, with the administrative and technical direction of the Library of Congress Network Development Office. It is one of a series of tasks in a project to provide information for a rational and cost-effective design for the development of a nationwide library data base.

Henriette D. Avram
Network Development Office
Library of Congress
January 1980

CONTENTS

LIST OF TABLES AND FIGURES iv

PREFACE vii

1. THE NATIONAL UNION CATALOG: OPERATIONS AND WORKFLOW 1

    1.1 Overview of Workflow and Volume of Work
        1.1.1 Organization of Operations
        1.1.2 Volume of Work
    1.2 NUC Control Section
        1.2.1 Catalog Maintenance Unit
        1.2.2 Searching Unit
    1.3 NUC Editorial Section

2. CHARACTERISTICS OF REPORTING LIBRARIES 11

    2.1 Types of Libraries and Volume of Reports
    2.2 Adherence to Reporting Guidelines
    2.3 Automated Capabilities of Reporting Libraries

3. DIFFERENCES IN HEADINGS BETWEEN OUTSIDE REPORTS TO THE NUC AND LC RECORDS 19

    3.1 Study Design
    3.2 Format for Subject Heading Analysis
    3.3 Differences in Subject Headings
    3.4 Format for Entry Heading Analysis
    3.5 Differences in Entry Headings
    3.6 Summary of Differences in Headings

4. DIFFERENCES IN BIBLIOGRAPHIC DESCRIPTION BETWEEN OUTSIDE REPORTS TO THE NUC AND LC RECORDS 31

    4.1 Format for Analysis of Description
    4.2 Differences in Description
    4.3 Comparison of Description and Heading Variations
    4.4 Time Taken to Replace NUC Records with LC Records

5. CONCLUSIONS 43

    5.1 Conclusions for a Nationwide Data Base Derived from the Analysis of Processes in the Catalog Publication Division
    5.2 Recommendations for Future Investigation

ABBREVIATIONS 51

LIST OF TABLES AND FIGURES

TABLE

1	Reports Received by Type of Library	12
2	Rate of Exclusions by Type of Library	14
3	Potential Contribution of Machine-Readable Reports	16
4	Distribution by Type of Library of Machine-Readable and Not-Machine-Readable Reports	17
5	Proportion of Reports from Each Type of Library That Are Potentially Available in Machine-Readable Form	18
6	Summary of Differences in Headings from Comparison of LC and NUC Records	28
7	Summary of Differences in Descriptive Elements from Comparison of LC and NUC Records	35

FIGURE

1	Decision Flow Chart of Macro Operations in the Production of the NUC	3
2	Distribution of Outside Reports	4
3	Subject Heading Differences: Cumulative Percentage Polygon	23
4	Entry Heading Differences: Cumulative Percentage Polygon	27
5	Total Number of Heading Differences per Record: Percentage Polygon	30
6	Bibliographic Description Differences: Cumulative Percentage Polygon	36
7	NUC Records Replaced by LC Records in 12-Month Intervals: Frequency Histogram	39
8	NUC Records Replaced by LC Records in 12-Month Intervals: Cumulative Percentage Polygon	41
9	Number of Records Published Annually in the NUC	45
10	Number of Records Added Annually to the OCLC System	46
11	Growth of the OCLC Data Base	47

PREFACE

The National Union Catalog was begun by the Library of Congress in 1901 with the stated purpose of becoming a central repository of bibliographic and location information concerning the resources of the major North American libraries. It was not, however, until January 1956 that the publication of union catalog information was begun with the conversion of the Library of Congress Catalog--Books: Authors into the National Union Catalog: A Cumulative Author List. At the time of first publication, only around 103,000 outside reports were being received annually for the catalog,1/ while in fiscal year 1978 over 4.5 million reports were received from libraries in the United States and Canada. The maintenance and publication of the National Union Catalog are currently under the control of the Catalog Publication Division of the Library of Congress. Since the publication of the original book catalog in 1956, the publishing function of the division has expanded to include the Register of Additional Locations; Subject Catalog; Films and Other Materials for Projection; Music, Books on Music, and Sound Recordings; Chinese Cooperative Catalog; Monographic Series; Newspapers in Microform; National Register of Microform Masters; Library of Congress Name Headings with References; and Symbols of American Libraries.

The National Union Catalog, the most extensive of the division's publications, has actively served the cause of scholarship and library effectiveness by providing current bibliographic and location information for the resources of American libraries and has materially contributed to the rational organization of acquisitions and resource sharing, cataloging, interlibrary loan, and reference and information work worldwide. The experience of the National Union Catalog provides information and guidance that are of immediate relevance to the design and planning of future nationwide data base and networking activities. Although it is presently a system based upon manual files and production procedures,2/ the fundamental characteristics of the system, including the establishment and continuous integration of records into a file created by decentralized input, are basic to union catalogs built by any type of procedure. This file (the catalog) has the integrity of a catalog of unique bibliographic records controlled by a single set of authorities and consistently applied national standards. Editorial procedures in the Catlaog Publication Division require that each record received from a participating library be subject to a critical examination and edit to ensure its conformance to the standards of a centralized Control File, including current Library of Congress practices and national bibliographic standards. Thus, the National Union Catalog presents itself as a working model that can be analyzed to project those activities required to fulfill some of the primary objectives for the library bibliographic component of a nationwide network.

The present study was undertaken to fulfill two principle objectives which would provide background data for the systematic development of the library bibliographic component of a nationwide data base and network. The first objective included analysis of the operations of the National Union Catalog in its current manual mode to determine some of the requirements for an automated nationwide union catalog. The second objective focused upon the characteristic differences of outside reports in comparison to Library of Congress records, in order to determine the extent of variation in bibliographic reporting that would probably exist in any future data base receiving decentralized input.

A mixed methodology was employed in gathering data which included examining summary statistics for fiscal year 1978; examining raw tallies of data for fiscal year 1978 kept for each reporting institution concerning number of reports and exclusions; interviews with key National Union Catalog personnel; observing searchers and editors in their work with representative searching and editorial problems; drawing a sample of 230 National Union Catalog reports replaced by Library of Congress records to ascertain length of elapsed time before replacement; and the comparison of valid lists or reporting libraries with published information indicating their capacity to deliver bibliographic records in machine-readable form. The specific methodology used with each set of findings will be detailed within the text of this report, when it is necessary to make certain assumptions explicit or to indicate limitations of the data gathered.

This project could not have been completed without the aid and encouragement of many people who generously contributed both ideas and information. I would like to express my gratitude to Gloria Hsia, chief of the Catalog Publication Division, for introducing me to the National Union Catalog and for thoroughly critiquing and improving the final manuscript; to John Reamy of the NUC Control Section and Barbara Petty of the NUC Publication Section for giving me access to current statistical information and guidance in observing the searching operations; to Frances Thackston, Starr Smith, and John Everete of the NUC Editorial Section for outlining editorial policies and operations and for permitting me to closely observe their editorial work; to David Graham and Jacqueline Lee, who served as research assistants in the most professional manner in spite of considerable pressures and obstacles; and to Sally McCallum, project monitor, whose guidance, encouragement, and affable criticism contributed measurably to the quality of the final product.

Raymond F. Vondran
Graduate Department of
Library and Information Science
The Catholic University of America
December 1979

NOTES

1. John W. Cronin, "History of the National Union Catalog, Pre-1956 Imprints," in Maurice F. Tauber and Hilda Feinberg, eds., Books Catalogs (Metuchen, N.J.: Scarecrow Press, 1971), p. 124.

2. Notable exceptions to these procedures are the machine processing of information appearing in the Register of Additional Locations and a small-scale experiment in the receipt of added locations in machine-readable form.

1. THE NATIONAL UNION CATALOG: OPERATIONS AND WORKFLOW

1.1 Overview of Workflow and Volume of Work

1.1.1 Organization of Operations

The Catalog Publication Division of the Library of Congress (LC) has the responsibility for the compilation and publication of the National Union Catalog (NUC) and related publications. The division is broken down into five major sections, which include the Office of the Chief, the NUC Control Section, the NUC Editorial Section, the NUC Publication Section, and the Special Catalogs Section. The focus of this report is upon the NUC Control Section and the NUC Editorial Section. The NUC Control Section has the major responsibility for the searching and disposition of outside reports and the maintenance of a Control File. This Control File contains all records that have been published in the NUC of post-1955 imprints and forms the primary control tool for that catalog. The NUC Editiorial Section edits outside reports for publication in the NUC, supplements the internal syndetic structure of the Control File, and ensures that duplicate records, entered under a different main entry element or different form, are not added to the Control File or published in the NUC.

The internal operations of the Catalog Publication Division are dependent upon searching and editing procedures applied to manual files. The exception to this procedure is the searching of the LC MARC data base when a submitted outside report is within the scope of LC MARC records. The limitations of coverage of the LC MARC file, comprising English-language imprints since 1968 and other Roman-alphabet languages being added between 1973 and 1977, restrict its present utility in daily operations. Thus, the process of searching the LC MARC file is followed neither frequently nor consistently due to the limitations on the scope of the LC MARC data, technical difficulties in searching the file, and the exigencies of the internal work flow.

The work flow begins with manual sorts to separate out those reports that are immediately identifiable as added locations for records already in the Control File, such as outside reports that carry an LC or an NUC card number on the record. At this stage records are also examined to determine whether they fall within reporting scope of the NUC guidelines for reporting. The remaining items are sorted alphabetically and searched in the Control File. After the initial examination and searching procedures have been accomplished, the work of the Editorial Section contributes to the integrity of the Control File and the published catalog by the further elimination of duplicate records and records excluded by the guidelines, by modifying outside reports to conform to the existing file structure and to current LC

practices, and by establishing the most judicious access point and references for records which will have limited access points within a book catalog. All records identified as duplicates of either LC records or previously published records are sent to the NUC Publication Section for incorporation into the Register of Additional Locations.

There are a series of natural constraints upon the overall organization of work and upon the general operational efficiency of the system that stem from the difficulty of manipulating catalog cards and accessing and maintaining a large manual file. Difficulties arise from searching a file the size of the Control File due to the large physical space required to house it, the complexity of alphabetical sequences, and the necessarily limited access points; from the physical inaccessibility of the LC Official Catalog, which is the final source of authority control for the Control File, due to the current remote location of the division; and from the absence of important reference tools for establishing entries and the necessary time to use these tools extensively if they were available yet still maintain the publishing schedule. A macro analysis of the major operations and decisions in the production of the NUC can be found in Figure 1.

1.1.2 Volume of Work

During fiscal year 1978 the Catalog Publication Division received 4,514,676 bibliographic records from 438 libraries in the United States and Canada. Although the mean number of reports per library was 10,307, there was wide variation in the extent of this reporting. Single reports were received from smaller institutions such as Graceland College, while large research libraries such as the University of Texas submitted as many as 146,849 records. Of the original outside reports submitted, about 2.8 percent (128,289) were immediately excluded from consideration because they did not fall within the NUC guidelines for reporting. On the basis of processing data reported for fiscal year 1978, 11.6 percent of the reports received were pre-1956 imprints, 56.1 percent were found to be known duplicates before bibliographic searching was undertaken, and 27.5 percent were found to be duplicates or outside of NUC reporting guidelines in the searching and editing process. Only 2 percent of the records submitted entered the NUC Control File as records for unique bibliographic items that will be published in the NUC of post-1955 imprints. The disposition of NUC reports is depicted graphically in Figure 2.

1.2 NUC Control Section

The NUC Control Section organizes the input into the National Union Catalog system. At this first stage in the operation all outside

FIGURE 1

DECISION FLOW CHART OF MACRO OPERATIONS IN THE PRODUCTION OF THE NUC

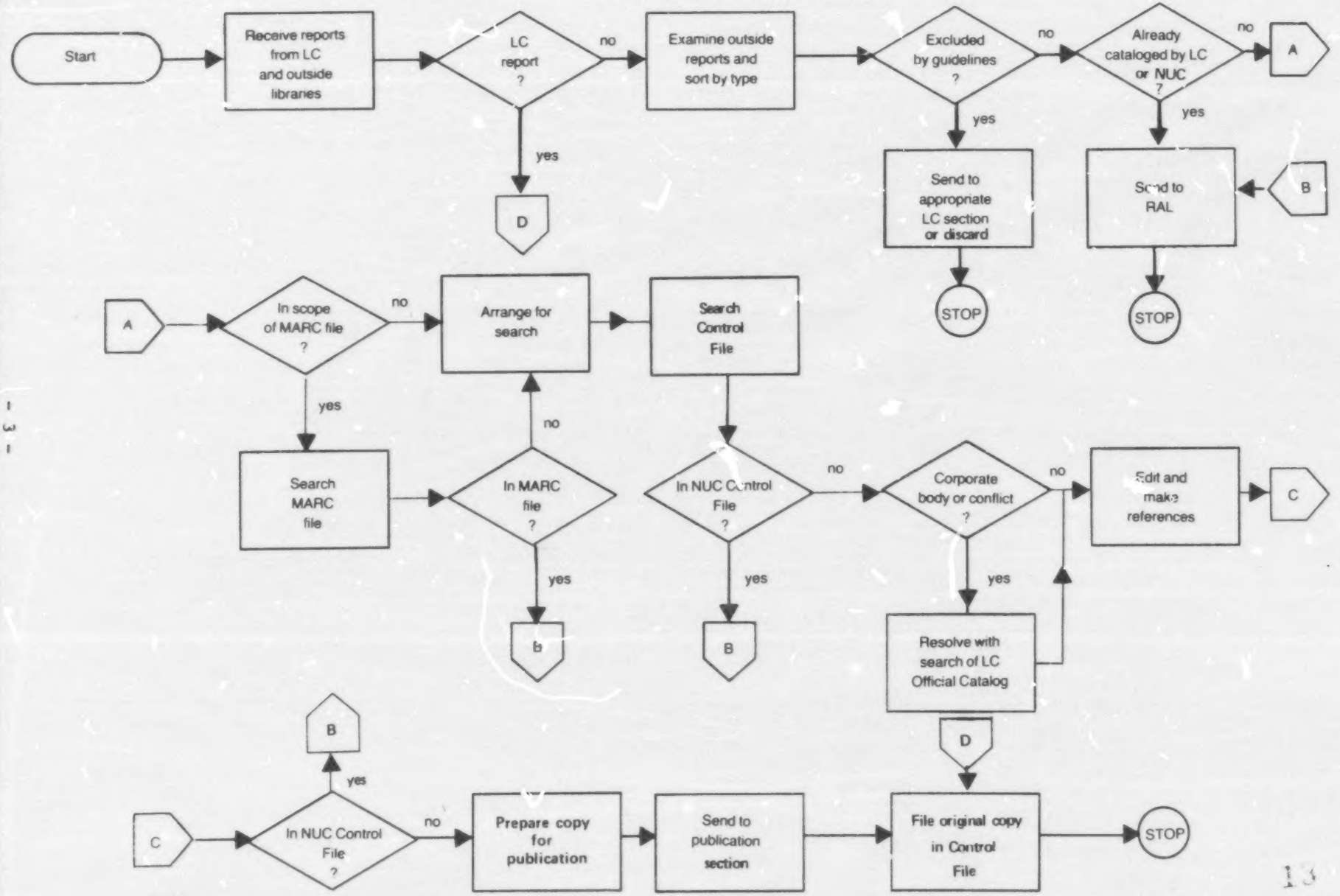
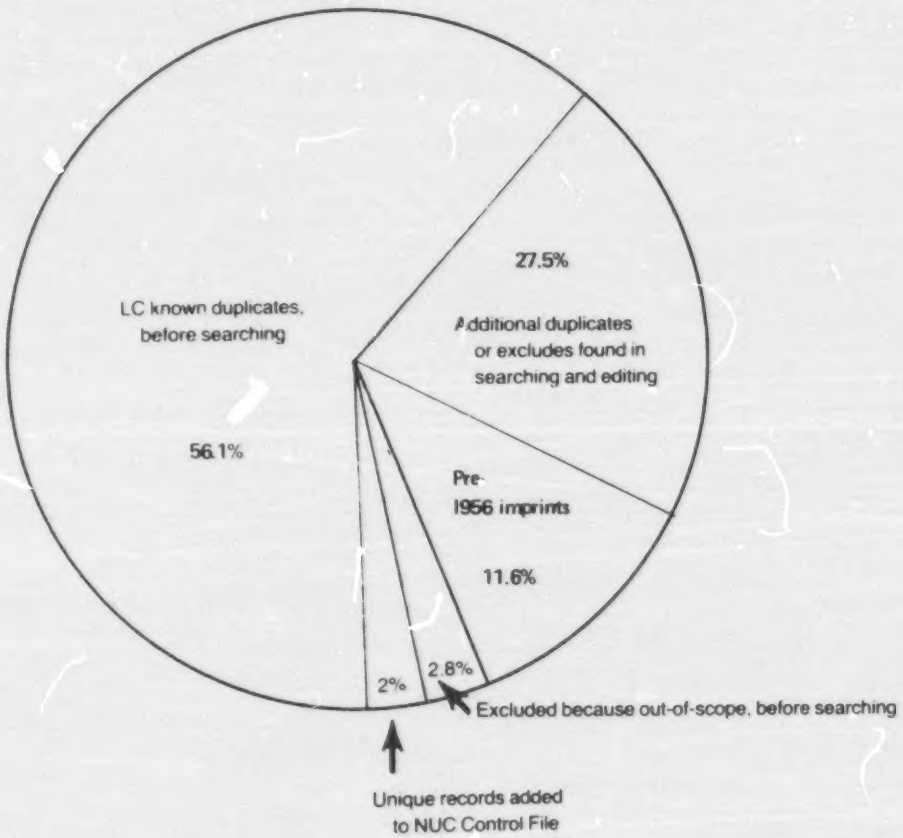


FIGURE 2
DISTRIBUTION OF OUTSIDE REPORTS
(Fiscal Year 1978)



reports are examined to determine whether they fit within the published guidelines established for full or selective reporting. Libraries participating in full reporting are asked to exclude from their reports the following: reprints, serials, United Nations publications, official state publications (with the exception of the one library in each state designated to report), and United States government publications (except for analytics in series not analyzed on LC cards). Selective reporting guidelines restrict libraries to submitting records for items that are foreign or outside the book trade, items for which LC cards are unavailable, and rare or unusual imprints. The Control Section administration contends that reporting libraries adhere closely to specified guidelines. The error rate, or the percentage of imprints received which are excluded by NUC guidelines, is relatively small: 2.8 percent of all reports received. To avoid unnecessary searches, all reports received are carefully examined for exclusions before searching.

Not all excluded items are indeed true 'errors' on the part of the reporting library. Some libraries, with the permission of the Catalog Publication Division, report all cataloged items regardless of guidelines, as a convenience to the reporting library and to ensure that all valid items are submitted for possible inclusion. Categories of exclusion are described with examples for searchers in the Searchers' Manual prepared by the Searching Unit of the NUC Catalog Control Section. Items regularly excluded include works in braille, broadsides, manuscript collections, honor's theses, domestic master's theses (except those in library science), microforms in series which have incomplete cataloging information, motion pictures, music scores, recorded interviews, pre-1956 imprints, seminar papers, orientalia, and serials. State and federal documents are entered into the Control File when they are new items, but subsequent reports of these items are not recorded as additional holding locations. Records for excluded items are forwarded, if appropriate, to other catalogs or units within the Catalog Publication Division or to other divisions of LC.

1.2.1 Catalog Maintenance Unit

The Catalog Maintenance Unit of the NUC Control Section maintains the Control File of 1956 and later imprints, the supplementary file to the NUC of pre-1956 imprints, and the file of pre-1956 imprints of the Slavic Union Catalog. The principle functions of the unit include filing in the post-1955 imprint Control File and signaling errors in filing, heading conflicts, and other problems in the syndetic structure of the file to unit supervisors, NUC editors, the Descriptive or Shared Cataloging Division, or the MARC Editorial Division. The NUC Control File which is maintained by the unit serves as a cumulative card repository of all entries found in the NUC of post-1955 imprints and as a source of authority for later items which may be established under the same or similar headings. Both the pre-1956 NUC and its supplement are

separate from the post-1955 NUC Control File, and non-LC headings established for outside entries are not necessarily checked and verified between the post-1955 and the pre-1956 files.

During fiscal year 1978 some 530,495 cards were added to the Control File, consisting of LC main entries, added entries, cross references, revised main entries, preliminary cataloging, and shelflist cards. If preliminary cataloging, references, added entries, and shelflist cards are excluded, cards for 291,341 unique bibliographic items were added to the Control File. Records for 68.3 percent of these items were received from LC and 31.7 percent were received from outside sources.

Although Catalog Publication is a division of LC and although most of the records contained in the Control File originated from LC, the variations in the headings of reporting libraries make it impossible to create an authority structure in the NUC Control File and the published catalogs that is identical with that of the LC Official Catalog. When outside reports are searched to remove duplication and the nonduplicates are forwarded to the editors for examination, they may be modified in both bibliographic description and entry headings to conform to the structures in the Control File (which is dominated by LC headings), LC policies and practices, the editorial practices of the division, and the Anglo-American Cataloging Rules (AACR). The principle sources of variation between the headings in the NUC Control File and the LC Official Catalog result from:

1) integration into the Control File of outside headings which have not been established in the LC Official Catalog;

2) integration into the Control File of outside headings which have been established in the LC Official Catalog, but, where, since no conflict for the heading existed in the Control File at the time of establishment and no search was made of the Official Catalog, the outside report was established as received;

3) the small number of cross references created for headings in the Control File. Because of publishing demands, not all recommended cross references are made in the Control File, as would have been the cost in the Official Catalog. (For example, inverted cross references are not made for conferences.)

Therefore, subsequent reports with the same or similar heading may be modified to conform to headings of outside reports already established in the Control File, which with its established reference structure serves as a form of authority. As a rule, however, when an LC record is entered into the Control File, that record and the form of its headings take precedence in the Control File.

If a heading has already been established in the Control File using a previous outside report, this heading will be maintained as an established heading if it conforms to the current cataloging rules and if there is no conflict. Only under special conditions, such as the establishment of a corporate entry, are outside headings searched in the LC Official Catalog before being established in the Control File. Thus, one can argue that the NUC Control File establishes an authority structure which is neither that of the LC Official Catalog nor that of individual libraries submitting reports but is instead that of a separate catalog. i.e., the Control File, receiving decentralized input.

1.2.2 Searching Unit

During fiscal year 1978, 701,478 records were searched in the NUC Control File and 1,370,402 records were received to be searched. As of September 29, 1978, there were 3,267,178 records waiting to be searched in the Control File, and there are indications that this backlog of records to be searched will not easily be diminished. Catalog Publication Division administrators report that the time taken for a record to pass through searching and be entered into the Control File ranges from a few days to twenty-two months. Despite the large backlog in searching, the average lapsed time for searching is less then one year. The items to be searched are in a temporary work file into which new reports are entered each day and from which searched reports are removed each day. Since searching is done in alphabetical sequence, however, how quickly a new report will be searched will depend on which letter is being searched when the new report enters the temporary work file.

Searches are conducted in the Control File by searchers with varying knowledge of cataloging rules and procedures. Cards are arranged in large alphabetical batches by main entry so that a single searcher may spend a considerable amount of time searching a few card drawers in close proximity. Searchers have specific searching quotas to meet, ranging from forty to fifty searches per hour or about 1,700 searches per week. As a result, searchers move quickly through record batches, entering the file under main entry and quickly comparing title and date fields as the most commonly used elements to determine duplicates. Secondary points used to discriminate between records are other imprint information, the series statement and the pagination. The edition statement is usually not as important an element for discrimination as the others mentioned. If a duplicate match is not found in the first pass, a searcher may later approach the Control File from an added entry, depending upon the information given in the author statement.

Searchers are generally not conversant with the AACR but follow the internally produced Searchers' Manual for determining duplicates.

If a matching record is not found in the file, a searcher will typically scan proximate records in the Control File to ensure that there has been no misfile of an existing record. If the first search does not result in a match, other clearly indicated entry points will be searched. Rather than attempt heuristic searches in the file, searchers will typically fill out a flag record in short form and place it into the Control File where the search has taken place and forward the record to the Editorial Section.

Because of the nature of duplicate searching, which is at one time labor intensive and also fixed rule oriented, attempts have been made in systems using machine readable data to write programs which will eliminate duplication in a file with little human intervention. A study was conducted for the Illinois State Library by Martha E. Williams to determine whether a state union catalog could be constructed with machine searches for internal duplication.1/ The objective of the study was to investigate an inexpensive technique for producing a file based upon decentralized input which would contain but one representation for each unique record. Two forms of error are inherent in any matching system. Mismatches occur when a unique and distinct record is thought to be duplicated in the file and eliminated as if it were a true duplicate. Missed matches occur when a true duplicate is added to a file and assumed to be a unique and distinct record. The philosophy of the investigators was that it was more detrimental to the system to err with mismatches because the intellectual content of a work was lost to the bibliographic system.

The primary search key for elimination of duplicates was the title field because it is a relatively good discriminator which appears in nearly all records and is least subject to variations in cataloging practice.2/ Other discriminators include date, pagination, and authorship. The tests were conducted using three files: 73,552 records (master file) from OCLC, Inc., 57,728 records from Northwestern University, and 22,857 records from the University of Chicago. Two passes were made of each record. On the first pass titles were compared, not character by character but by a bit string comparison called a Harrison Key, in which variations in titles due to transcription and clerical error were reduced by applying a Hamming distance criteria for matching which permitted some title variation between duplicate records. Date, pagination, and finally authorship were used in the second pass for a higher level of discrimination. This sequence for identifying duplicates (title, date, pagination) is similar to the search strategies observed at the NUC in the conduct of this investigation.

Test results indicate this machine strategy to be highly effective in eliminating duplication from a file. Tests using the title/date key alone indicate a matching effectiveness of 97.10 percent.3/ Using additional combinations, error was reduced to 0.0704 percent on matches of 2,838 pairs of records. Using title/date/LC card

number/ISBN search keys the effectiveness of the procedure was 98.62 percent.4/ Through these procedures only 1.1 percent of the records required manual verification. Such procedures could be used for removing duplicates and signaling locations in a national union catalog or nationwide data base that was based upon machine-readable records and operations.

1.3 NUC Editorial Section

The NUC Editorial Section is responsible for integrating records from outside sources into the Control File of the NUC. The work of the editorial staff is both heuristic and at times creative, attempting to ensure that duplicate records are not added to the Control File, modifying records to conform to the established file structure and authority, and applying the AACR as adopted under LC's policy of superimposition. Editors correct and modify both entry headings and descriptive cataloging of outside reports but make no changes in subject headings or classification. The Control File serves as the primary authority for integrating records. The LC Official Catalog is only searched by assistant editors when there is a conflict between headings in the Control File or when an outside report contains a corporate main entry that has not been established in the Control File, a personal name of nobility, or the name of an author who flourished before the nineteenth century.

The major variation between NUC editorial practice and LC practice is in the interpretation of AACR concerning added entries and in the quantity of cross references made. This variation results from the editorial demands of publishing a book catalog and the difficulty in integrating records without the items that they describe in hand. In editing outside reports for inclusion in the Control File, editors are at liberty to choose the best and fullest record that conforms to LC practices and to the AACR. If additional added entries not found on the record are required, however, additional access points are not made. Thus, if an outside report contained two author added entries and a title added entry, and these were within LC practice and specified by the AACR, these additional access points would be made. On the other hand, if an identical record were to be submitted without these added entries, these additional access points might not be made, even if the bibliographic description clearly indicated that application of the AACR would require additional added entries, because it is difficult to be certain of the entries without the item being cataloged in hand.

The resources for editing include the AACR, the Control File, NUC editorial policy, the internal and published policy decisions of the LC cataloging divisions, conversations with members of the LC Descriptive Cataloging Division, the LC Official Catalog, the recommendations of

individual libraries considered authorities in specific areas, and the most likely approach of the user of the published NUC. Constraints upon the system include the limited access Control File, the requirements of a published book catalog, the restrictive crossreference structure of both the Control File and the requirements of space in a publication, and the inherent problems of modifying a cataloging record without the document in hand. While written internal NUC editing guidelines are available to editors, difficulties in keeping them current has diminished their usefulness. A recent updating of a specific subset of the guidelines in the area of Catholic Church and Church History has resulted in greater editorial consistency, according to the editorial staff. The extension of detailed editing guidelines into other areas would perhaps produce the same results. Because of these constraints, many of the editorial decisions must be based upon the individual judgment and the experience of the editor and the editor's knowledge of practices at and quality of the cataloging records of specific institutions, while taking into account the limitations of a file and publication designed for main and added entry access.

There is a tendency on the part of the editorial staff to choose title entry when confronted with a variety of practices, providing added entries from other approaches as possible. The heading on an outside report is modified when the heading has been established by an LC record or a previous outside report. Changes in choice or form of heading will also be made when a conflict exists in the Control File or information on the record obviously varies from AACR standards.

One means of creating a more consistent Control File that would serve as a base of high integrity as a national file would be to have greater access to authorities of LC and to create a written set of editorial guidelines which could also be distributed to libraries who supply reports to the NUC. The inherent limitations of manual files and the space limitations of a publication with limited access points will continue to prevent the sharing of bibliographic information on a national level to reach its potential for efficient access to a data base under a uniform system of authorities.

REFERENCES

1. Martha E. Williams and Keith D. MacLaury, A State-Wide Union Catalog Feasibility Study, Final Report on Project III-FY 1976 for the Illinois State Library (Urbana, Ill.: Information Retrieval Research Lab, University of Illinois, 1977).

2. Ibid., p. 28.

3. Ibid., p. 82.

4. Ibid.

2. CHARACTERISTICS OF REPORTING LIBRARIES

2.1 Types of Libraries and Volume of Reports

The 438 libraries that submit reports to the Catalog Publication Division for the NUC break down into six basic types according to their specialization or subject collection strengths: business, theology, humanities and social science, law, scientific and medical, and general collections. Each of the reporting libraries was classified into one of these mutually exclusive and exhaustive categories by consulting the American Library Directory to determine specific subject strengths. The number of reports received by the Catalog Publication Division during fiscal year 1978 was then calculated for each type of library reporting.

The vast majority of outside reports (about 86.4 percent) were received from 51 percent of the libraries which fell into the category of libraries containing general collections--the larger U.S. academic and research libraries. A relatively small portion of the major university and research libraries provided a large proportion of the total reports received for the NUC in fiscal year 1978. Twenty-three major libraries, which represent only 5.25 percent of the libraries reporting, contributed 34 percent of the records submitted to the NUC. Although the plan for data gathering had not originally included the category of theology libraries, it was found in the data collection to represent 11 percent of all libraries reporting and surpassed both law and business libraries in terms of the number of items reported. Table 1 lists the proportion of reports submitted by each type of library and the proportion of reporting libraries classified in each type.

Turning from the analysis of the number of reports received to the number of new unique records that are actually added to the NUC Control File, the same pattern is again of large research libraries making the greatest contribution of records. In a survey conducted by the Library of Congress staff, a systematic sample was taken of the non-LC reports listed in the four volumes of the July-September 1976 quarterly edition of the NUC. This sample consisted of 108 clusters of seven records each and yielded 754 unique non-LC records which were analyzed in terms of the number of records contributed by each library. These libraries were categorized into the basic six types of libraries by collection strength. Again, a small number of predominently research and university libraries contributed a high percentage of non-LC records for the publication. Twenty-three libraries, which represented only 12 percent of the 191 libraries in the sample, provided approximately 42.6 percent of the unique non-LC records.

It would be easy to conclude that the main contributors to a future nationwide data base would be the major American university and research libraries. Such extrapolation from the data derived from study

TABLE 1

REPORTS RECEIVED BY TYPE OF LIBRARY
(Fiscal Year 1978)

TYPE OF LIBRARY	REPORTS RECEIVED		LIBRARIES	
	No.	%	No.	%
Business	10,492	.2	4	1
Theology	98,233	2.2	49	11
Humanities/Social Science	169,128	3.7	67	15
Law	77,251	1.7	21	5
Science/Medicine	255,234	5.7	72	16
General Collections	3,903,252	86.4	225	51
TOTAL	4,514,676	99.9	438	99

of the NUC reporting must be made with due caution, however. Since both reporting and using bibliographic information are in part functions of perceived cost and benefit to the individual libraries who contribute to and use the NUC, participation may be limited by the following factors: the difficulties in reporting, the costs of access to the NUC, the obsolescence of the bibliographic data, the use of alternative systems such as OCLC, Inc., Research Libraries Information Network (RLIN), or Washington Library Network (WLN), and the inflexibility of main entry cataloging information bound by a single set of standards. These may tend to limit participation in specific ways that would not be the same in a nationwide on-line system which offered each library a different set of cost/benefit considerations.

2.2 Adherence to Reporting Guidelines

There is little difference in the rate at which reports from various types of libraries are excluded by NUC reporting guidelines, with the exception of the humanities and social science libraries, which have a rate of contributed out-of-scope reports to in-scope reports of 2:1. Libraries within the humanities and social science group are generally smaller special libraries, museum and society libraries, or university departmental libraries. It may be more difficult to enforce guidelines upon the small independent libraries of this group, but the special collections of these institutions may make this rate of out-of-scope reports tolerable when compared to their overall contribution of potentially unique holdings. The rate of exclusion by type of library is displayed in Table 2.

2.3 Automated Capabilities of Reporting Libraries

One of the most crucial questions in anticipating the requirements for the development of a nationwide data base with on-line access is the extent to which potential participants have the capacity to transmit records in machine-readable form. The strategy in this study was to use the NUC libraries as a core of potential participants in this data base and to assess their automation capability. Data were gathered to answer this question by examining NUC statistics on reporting libraries and producing a list of all libraries that had submitted reports to the NUC during fiscal year 1978. This list was searched in the published literature to determine which libraries were either active members of an on-line network or information utility, or had an operational in-house capability of producing and maintaining records in machine-readable form. The number of records potentially available in machine-readable form was then calculated on the basis of NUC statistics on the number of records submitted by each reporting library during fiscal year 1978. These data are subject to one

TABLE 2

RATE OF EXCLUSIONS BY TYPE OF LIBRARY
(Fiscal Year 1978)

TYPE OF LIBRARY	OUT-OF-SCOPE RATE	OUT-OF-SCOPE REPORTS (%)	CONTRIBUTED REPORTS (%)
Business	1.9	.15	.2
Theology	3.4	2.58	2.2
Humanities/Social Science	5.6	7.58	3.7
Law	2.6	1.59	1.7
Science/Medicine	2.4	4.70	5.7
General Collections	2.7	83.57	86.4
TOTAL (Percent)		99.91	99.9
TOTAL (Number)	2.8	128,289	4,514,676

important unverified assumption: libraries that have a machine-readable bibliographic capability actually use this capability for all bibliographic records they submit to the NUC. A secondary assumption is that all libraries that have this capability are known to this investigator.

Of the 4,514,676 outside reports received by the NUC during fiscal year 1978, 89.3 percent were potentially available in machine-readable form (Table 3). As might be expected in the light of their large contribution of reports to the NUC (86.4 percent), the largest potential contributors of machine-readable records were the libraries with general collections (research and academic) which were capable of supplying 90.5 percent of the machine-readable reports. The other types of libraries taken together would only be able to provide 9.5 percent of all machine-readable reports--less than their overall present contribution of reports to the NUC (13.6 percent). Thus it can be noted from Table 4 that the proportion of reports would diminish from all other types of libraries if input and access were only available in machine-readable form.

An examination of Table 5 shows that the majority of reports contributed by each type of library were potentially available in machinereadable form and that a majority of libraries in business and theology as well as libraries with general collections had this capability. Thus it would seem to be feasible for a nationwide network to orient its operations to machine-readable rather than manual procedures for bibliographic reporting. This would accommodate the majority of libraries potentially reporting and would orient operations to the vast majority of all records received.

Moreover, if we consider the three major bibliographic utilities' contribution to the potential national library network data store, OCLC, RLIN, and WLN would be able to report 95.9 percent of all machine-readable records available from NUC participating libraries. This would include the reports of 97.7 percent of all libraries that have machine-readable bibliographic capabilities. Lastly, linkage with these three utilities would potentially provide access to 85.6 percent of all reports received for the NUC during fiscal year 1978.

TABLE 3

POTENTIAL CONTRIBUTION OF MACHINE-READABLE REPORTS
(Fiscal Year 1978)

REPORTS/LIBRARIES	NUMBER	PERCENT
Reports available in machine-readable form	4,032,108	89
Reports not available in machine-readable form	482,568	11
TOTAL REPORTS	4,514,676	100
Libraries capable of reporting in machine-readable form	268	61
Libraries not capable of reporting in machine-readable form	170	39
TOTAL LIBRARIES	438	100

TABLE 4

DISTRIBUTION BY TYPE OF LIBRARY OF
MACHINE-READABLE AND NOT-MACHINE-READABLE REPORTS
(Fiscal Year 1978)

TYPE OF LIBRARY	COULD REPORT IN MACHINE-READABLE FORM		COULD NOT REPORT IN MACHINE-READABLE FORM		TOTAL	
	Reports(%)	Libraries(%)	Reports(%)	Libraries(%)	Reports(%)	Libraries(%)
Business	.2	1	.5	1	.2	1
Theology	1.5	10	7.4	14	2.2	11
Humanities/Social Science	2.7	8	11.9	27	3.7	15
Law	1.0	4	7.3	6	1.7	5
Science/Medicine	3.9	11	20.6	25	5.7	16
General Collections	90.5	66	52.3	27	86.4	51
TOTAL (Percent)	99.8	100	100.0	100	99.9	99
TOTAL (Number)	4,032,108	262	482,568	170	4,514,676	438

TABLE 5

PROPORTION OF REPORTS FROM EACH TYPE OF LIBRARY THAT ARE POTENTIALLY AVAILABLE IN MACHINE-READABLE FORM
(Fiscal Year 1978)

TYPE OF LIBRARY	COULD REPORT IN MACHINE-READABLE FORM		COULD NOT REPORT IN MACHINE-READABLE FORM		TOTAL	
	Records(%)	Libraries(%)	Records(%)	Libraries(%)	Records(No.)	Libraries(No.)
Business	77	50	23	50	10,492	4
Theology	63.7	53	36.3	47	98,233	49
Humanities/Social Science	65.9	31.3	34.1	68.7	169,128	67
Law	55	47.6	45	52.4	77,251	21
Sciences/Medicine	61.2	41.7	38.8	58.3	255,234	72
General Collections	93.5	79.6	6.5	20.4	3,903,252	225

3. DIFFERENCES IN HEADINGS BETWEEN OUTSIDE REPORTS TO THE NUC AND LC RECORDS

The analysis of the extent of differences in both entry and subject headings between outside library reports added to the NUC Control File of the Catalog Publication Division and those LC records which later replaced these reports gives us an opportunity to observe difficulties that will be encountered in bringing together records created by various bibliographic standards and practices in establishing a nationwide data base. It is expected that the cataloging created in different institutions will vary due to a number of specific factors, including the reference sources available to the cataloger, the publications for which the required headings were first established, the cataloging codes used, the cataloging guidelines and policies established in individual institutions, the file structure of the existing catalog into which new records are incorporated, the individuals who interpret the rules, and and the size of the collection and the catalog of the institution. The integrity of a national file of bibliographic information based upon decentralized input will depend upon a consistent and uniform policy for establishing headings. In a sense, examination of differences between LC and outside headings is a comparison between different bases of bibliographic data, under separate but interrelated authority systems and policy guidelines.

3.1 Study Design

A sample of bibliographic records was taken to reliably estimate the extent of variations between LC cataloging practices and the practices of libraries that reported their holdings to the NUC. This sample study was broadly modeled after a study comparing outside reports and LC replacements on specific data elements carried out by Gloria Hsia, chief of the Catalog Publication Division, in 1969. That study examined a small sample of replaced outside reports within the broad categories of entry headings, subject headings, and bibliographic description. The present study limits itself to variations in both entry and subject headings found on a sample of outside reports which had been added to the Control File and published in the NUC and those found on the LC records that replaced the outside reports. The headings are compared for specific, detailed categories of differences. The categories used in the analysis of these differences were designed to be consistent with other tasks directed by the Network Development Office so that meaningful comparisons could result between various studies. It should be noted that the copies of outside library reports that are added to the Control File are the original records received from the reporting libraries, with editing changes marked, not the retyped version that appears in the printed NUC. The comparisons were made between the data as provided by the reporting library and the

corresponding LC records produced later (and independently) by the LC cataloging divisions.

A sample design was developed to determine the extent of differences between NUC outside reports and the LC records that replaced them at a level of 90 percent confidence and 5 percent tolerance. Thus, there would be a 90 percent chance that proportions calculated from the data in the sample would be within 5 percent of the true population proportion. To determine the smallest sample size that would be sufficient to meet these requirements, a measure of the homogeneity of the population on the variable deemed most critical (the number of reports that had at least one variation in an entry heading when compared to the LC record that replaced it) was required. The best available estimate of this homogeneity was provided by the study conducted by Mrs. Hsia. Using this measure of homogeneity as a correction factor, it was determined that a sample of 250 outside reports replaced by LC records would be adequate to achieve the required level of confidence and tolerance. Post hoc analysis revealed that a sample size of 233 would have been adequate to achieve the purposes of the design.

A simple random sample was taken from the 1978/79 outside reports replaced by LC records that were held by the Searching Unit of the Catalog Publication Division. An outside library report may survive the screening for duplication and exclusion and be added as a unique record to the Control File and published in a monthly or quarterly volume of the NUC but then be superseded by an LC record. Superseded records are held for a period by the Searching Unit. The 250 record sample was copied and each record was searched in the Control File and compared in terms of variations in entry and subject heading access points with the LC record that replaced them. Differences were noted on each copy of the outside report and coded by the principle investigator. A trial experiment to determine the reliability of coding was undertaken before the test sample was drawn and analyzed. The specific categories of differences and the results of the data analysis are given in the following sections.

3.2 Format for Subject Heading Analysis

The categories for analyzing subject heading differences between outside reports and LC replacements were developed to be consistent and comparable with other research concerned with developing general guidelines for applying LC subject headings in a network context, as outlined in Task 11 of the Nationwide Data Base Design Project.1/ Specifically, these categories were created for use in the analysis of the variations between LC subject heading practices and records in the Northwestern University Library Africana file and the serial records in the Conversion of Serials (CONSER) project. The perspective of these

categories of comparison is from that of the LC record. The categories used for subject heading analysis were:

- o Additional LC Subject Heading
- o Absence of Outside Library Subject Heading
- o Variation in Form of Subject Heading
- o Additional LC Subject Subdivision
  - form subdivision
  - chronological subdivision
  - geographic subdivision
- o Absence of Outside Library Subject Subdivision
  - form subdivision
  - chronological subdivision
  - geographic subdivision
- o Variation in Form of Subject Subdivision
  - form subdivision
  - chronological subdivision
  - geographic subdivision

An Additional LC Subject Heading was defined as the presence of a distinct subject heading on a record created by LC where no such distinct heading was present on the outside report. An Absence of Outside Library Subject Heading was defined as the absence of a distinct subject heading on the LC record where one existed on the outside library report. Variation in Form of Subject Heading was defined as a difference in the form or the fullness of form of the same subject heading that appeared on both an outside report and the LC record that replaced it.

3.3 Differences in Subject Headings

Considerable differences were found between the subject headings of outside reports and the subject headings of the LC records that replaced them. Of the outside reports, 73.2 percent displayed at least one subject heading difference when compared to the subject headings provided in the LC record.

(1) Additional LC Subject Heading:

The greatest source of difference between subject headings was due to the LC record providing a subject heading not contained on the outside report. These constituted 51.2 percent of all subject heading differences.

(2) Absence of Outside Library Subject Heading:

The area of second greatest difference occurred in the absence of a subject heading on the LC record that was present on the

outside report. These constituted 28.8 percent of all subject heading differences.

(3) Variation in Form of Subject Heading:

Only 2.6 percent of subject heading differences occurred because of variations in the form of subject heading between the LC record and an outside report.

(4) Subject Subdivision Differences:

Of all subject heading differences between the records compared, 17.4 percent were due to some specific difference in a subject subdivision where the main subject entry element remained the same. These difference were distributed as follows:

Additional LC Subject Subdivision: 9.4 percent
- form subdivision: 6.4 percent
- chronological subdivision: 0.2 percent
- geographic subdivision: 2.8 percent

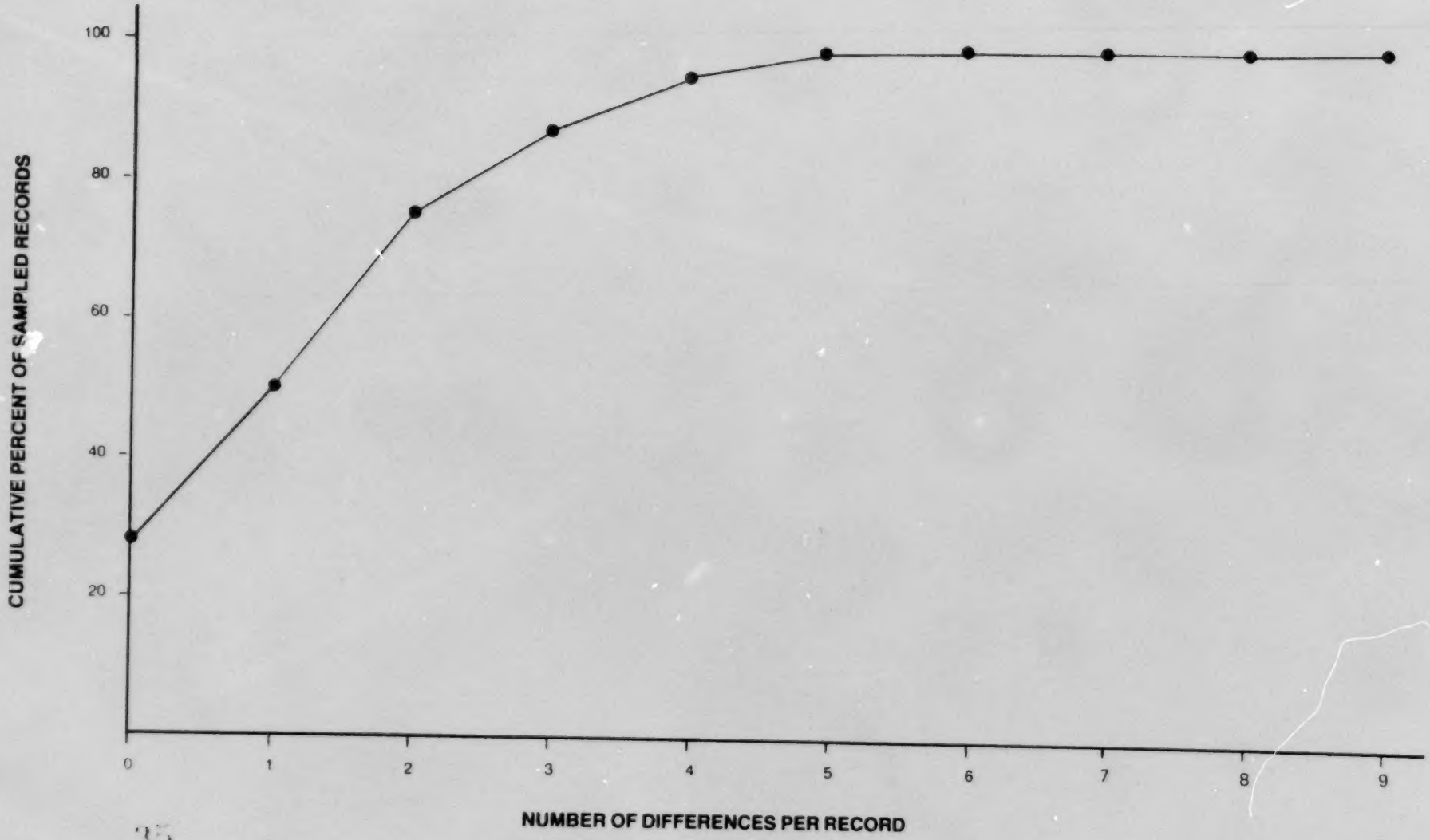
Absence of Outside Library Subject Subdivision: 1.9 percent
- form subdivision: 1.2 percent
- chronological subdivision: 0.2 percent
- geographic subdivision: 0.5 percent

Variation in Form of Subject Subdivision: 6.1 percent
- form subdivision: 3.9 percent
- chronological subdivision: 0.5 percent
- geographic subdivision: 1.7 percent

The information provided in Figure 3 shows the cumulative percentage of records in the sample having specific numbers of subject heading differences. It can be seen from the figure that almost half (49.6 percent) contain one subject heading difference or less. Of the records compared, 26.8 percent had no differences. The mean number of subject heading differences per record compared was approximately 1.7.

Similar results were reported for analysis of subject headings assigned to serial records by participants in the CONSER project. In a sample of 99 cataloging records which had been originally cataloged by CONSER members and then used by LC for current cataloging, 45.9 percent of all differences resulted from Additional LC Subject Headings, in contrast to the NUC sample which indicated that 51.2 percent of all differences had resulted from additional headings. Some differences in the results, which could be accounted for by either sampling procedures or differences in the nature of cataloging serials as opposed to monographs, can be noted in the comparison of the two studies. The serial sample showed that 43.2 percent of the differences resulted from the Absence of Outside Library Subject Headings, while the NUC sample indicated that only 28.8 percent of the differences were due to this factor. Moreover, the study of CONSER records indicated that 10.8

FIGURE 3
SUBJECT HEADING DIFFERENCES:
CUMULATIVE PERCENTAGE POLYGON
(N=250)



percent of all differences resulted from Variations in Form of Subject Heading or some difference in subject subdivisions, while analysis of the NUC data resulted in 20 percent of all differences being attributed to differences in these categories.

3.4 Format for Entry Heading Analysis

Differences in entry headings were compared between the 250 sampled outside library reports that were added to the Control File and the LC records that replaced them. The categories developed for analyzing these differences were created to satisfy three distinct requirements: to make the findings on entry headings as comparable as possible with those findings on subject headings in this and other research; to make them consistent with the major classifications of headings in the AACR; and to highlight the impact of specific differences in headings when maintaining a limited-access file, such as the NUC Control File. As was true in the analysis of subject heading differences, the perspective of these categories of comparison was from that of the LC record. The specific categories for entry heading analysis were:

- o Differences in Main Entry Headings
  - choice of entry
  - form of entry
- o Additional LC Added Entry Heading
  - author (including corporate author)
  - title (including uniform and alternative titles)
  - series
- o Absence of Outside Library Added Entry Heading
  - author
  - title
  - series
- o Variation in Form of Added Entry Heading
  - author
  - title
  - series

3.5 Differences in Entry Headings

The analysis of the differences in the sample indicate that approximately 68 percent of all records had at least one heading difference when an LC record replaced an outside library record. In this calculation of differences, the convention of a main entry was subscribed to, because the NUC Control File and the printed NUC provided main entry access. Thus, two differences would be scored for a record when the outside report had included a main and added entry in the

reverse order to that found on the LC record. This occurred in 11 of the 250 records sampled. But even if we abandon the notion of a main entry access point and treat all headings as equal access points, 63.2 percent of all records compared had at least one difference in entry headings. The specific findings of the comparison of records is detailed below.

(1) Additional LC Added Entry Heading:

The largest percentage of differences in records was due to the LC record having a distinct added entry heading when no such distinct heading was present on the outside report. This case constituted 50.2 percent of all differences in entry headings. Of all the Additional LC Added Entry Headings:

15.2 percent were due to additional author headings
20.0 percent were due to additional title headings
14.3 percent were due to additional series headings

(2) Absence of Outside Library Added Entry Heading:

The second largest percentage of difference was due to the absence of outside added entry headings on LC records, which represented 21.9 percent of all differences. These include:

12.4 percent of differences due to the absence of author headings
6.3 percent of differences due to the absence of title headings
3.2 percent of differences due to the absence of series headings

(3) Differences in Main Entry Headings:

The third largest number of entry heading variations (19.6 percent) were due to differences in the main entry between the two records compared in either choice or form of entry. Of the differences, 7.6 percent were due to the choice of main entry and 12.0 percent were due to form of the main entry. About one-third of the differences in form of main entry were due to variations in the fullness of form. In eight cases the LC record actually diminished the fullness of form, which could either be a result of file characteristics or of the application of the newer title-page-oriented rules. In an equal number of cases LC increased the fullness of form, perhaps again due to its own file characteristics. In either case, these differences had no bearing on resolving conflicts within the NUC Control File.

(4) Variation in Form of Added Entry Heading:

Variations in the form of added entries accounted for 8.5 percent of all differences detected between outside reports and LC records. Of these, 2.9 percent were variations in the form of name entry, 1.2 percent were variations in the form of title, and 4.4 percent were due to variations in the form of series added entry. However, if the convention of a main entry is abandoned and the differences in the form of those entries are added to this total, the total number of differences due to variations in form increases to third place among the major sources of differences and represents 20.6 percent of the total number of entry heading differences.

Figure 4 displays the cumulative percentage of records in the sample in relation to the number of entry heading differences found in each record. It can be seen that approximately 66 percent of all records contained one entry heading difference or less. Because the distribution is skewed and there are a number of extreme scores, the mean calculated at 1.26 differences per record is not an appropriate "average" value. Since both the median and the mode are the same, it would be most useful to characterize the average number of entry differences anticipated through replacements as one.

3.6 Summary of Differences in Headings

The results of the analysis of the NUC sample data showed that 42.6 percent of all differences in outside reports were due to differences in entry headings, while 57.4 percent were due to differences in subject heading entry elements.

The majority of differences that result from replacement of outside library reports by LC records are additional LC entry and subject heading access points. Of all record changes, 50.8 percent result from additional LC headings, with a split of 21.4 percent in entry headings and 29.4 percent in subject headings. By contrast, 25.8 percent of all changes were due to the absence of headings from outside library headings, 9.3 percent in the absence of entry headings and 16.5 percent in the absence of subject headings. All other differences, accounting for 23.3 percent of the total, are the result of minor categorical differences which may be observed in Table 6.

There appears to be consistent evidence to support the conclusion that when an LC record replaces an NUC outside report, additional points of access are created. Of the total of 739 bibliographic variations made in the sample of 250 records, 50.7 percent of all differences were due to additional entry or subject access points in LC records (375), while 25.8 percent of all differences (191) were due to the absence of

FIGURE 4

ENTRY HEADING DIFFERENCES: CUMULATIVE PERCENTAGE POLYGON (N=250)

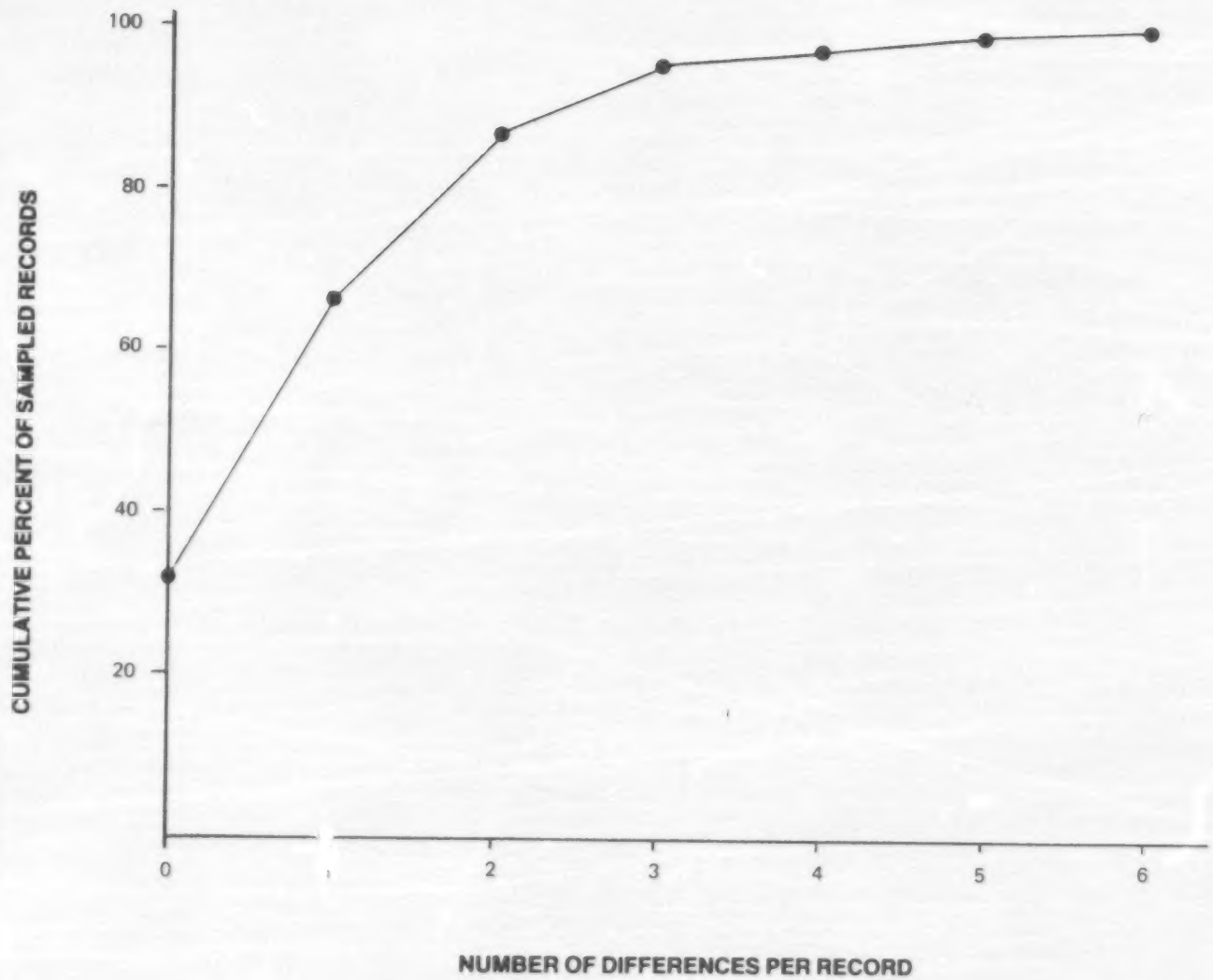


TABLE 6

SUMMARY OF DIFFERENCES IN HEADINGS FROM
COMPARISON OF LC AND NUC RECORDS
(N = 250)

TYPE OF DIFFERENCE	NUMBER OF DIFFERENCES		PERCENT OF TOTAL DIFFERENCES	
Additional LC Headings	375		50.8	
Additional LC Entry Heading		158		21.4
Additional LC Subject Heading		217		29.4
Absence of Outside Library Headings	191		25.8	
Absence of Outside Library Entry Heading		69		9.3
Absence of Outside Library Subject Heading		122		16.5
Other Differences	173		23.3	
Differences in Main Entry Heading, Form		38		5.1
Differences in Main Entry Heading, Choice		24		3.2
Variation in Form of Added Entry Heading				
- name		9		1.2
- title		4		.5
- series		14		1.9
Variations in Form of Subject Heading		11		1.5
Variations in Form of Subject Subdivisions				
- form and topical subdivisions		16		2.2
- chronological subdivision		2		.3
- geographic subdivision		7		.9
Additional LC Subject Subdivision				
- form and topical subdivisions		27		3.7
- chronological subdivision		1		.1
- geographic subdivision		12		1.6
Absence of Outside Library Subject Subdivision				
- form and topical subdivision		5		.7
- chronological subdivision		1		.1
- geographic subdivision		2		.3
TOTAL	739	739	99.9	99.9

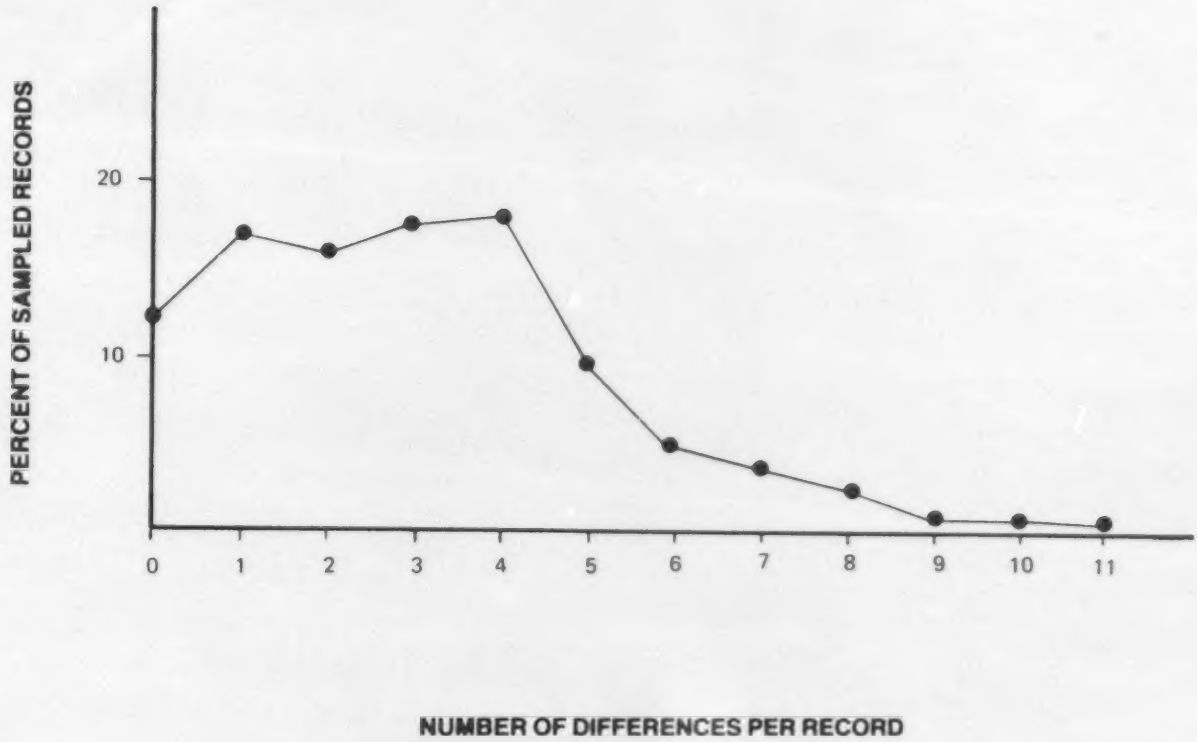
entry or subject access points on the LC record. When additions and absences are compared, an average (mean) of .42 additional access points per record were made in the LC record. The ratio of additions to absences of subject headings was 1.8:1. The overall ratio of additions to absences of access points was 1.96:1. The LC replacement of NUC outside reports resulted in a 1.2 percent increase in name access points, a 5.8 percent increase in title access points, a 4.7 percent increase in series access points, and a 12.9 percent increase in subject access points.

Both the mean, 2.96, and the median, 3.0, figures indicate that the approximate "average" number of differences that can be anticipated in any bibliographic record is three. Figure 5 displays the distribution of differences in each record in relation to the total percentage of differences. If one wished to predict the number of differences that will occur per record, a good estimate would be between one and four differences per bibliographic record.

REFERENCES

1. Edwin J. Buchinski, Initial Considerations for a Nationwide Data Base, Network Planning Paper no. 3 (Washington: Network Development Office, Library of Congress, 1978), pp. 34-36.

FIGURE 5
TOTAL NUMBER OF HEADING DIFFERENCES PER RECORD: PERCENTAGE POLYGON (N=250)



4. DIFFERENCES IN BIBLIOGRAPHIC DESCRIPTION BETWEEN OUTSIDE REPORTS TO THE NUC AND LC RECORDS

4.1 Format for Analysis of Description

Differences in descriptive cataloging were analyzed between the 250 records in the sample of outside library reports that entered the NUC Control File and the LC records that replaced them. The sample was that described in section 3.1. The categories for analysis were developed to parallel the major elements of description developed in the AACR. Differences were coded with the historical objective of descriptive cataloging that has been developed since Cutter's Rules for a Dictionary Catalog: "To state the significant features of an item with the purpose of distinguishing it from other items and describing its scope, contents, and bibliographic relation to other items."1/ Thus, differences were noted when the specific variation would either affect the identification of the item. Those that affected neither could be reasonable ignored.2/ Differences in internal punctuation which resulted from the introduction of the International Standard Bibliographic Description (ISBD), e.g., hyphenation between description fields and slashes between title and author statements, were ignored. Minor differences in abbreviations (e.g., "ill." for "illus.") were also not scored as variations. Because nearly all records differed on the specification of size found in the collation, which could result from differences in binding standards, among other variables, any variations found in this area were not scored as differences.

The specific categories for the analysis of descriptive differences and the types of modifications each included were:

- Variations in Title Statement

  These include modification of the title or subtitle, increased or diminished completeness, additions or absences of uniform, alternate, or parallel titles, etc.

- Variations in the Author Statement

  These include increased or diminished completeness of this statement, such as addition or absence of joint authors, editors, translators, etc. Nonsubstantive words such as "by", "Hrsg. von", etc., were ignored.

- Variations in the Edition Statement

  These included any substantive difference in the statement such as numbering and distinguishing phrases like "revised," "enlarged," etc.

- Variations in Place and Publisher

  These include substantitive changes in places and publishers' names such as the addition of place names, distributors, etc. Variations in the abbreviation of these names were ignored.

- Variations in Date of Publication

  These include any difference in date such as addition or absence of copyright date, printing date, etc. Differences in punctuation were ignored.

- Variations in Pages or Volumes

  These include any difference in paging such as numbering of preliminary pages, numbering of plates, and the addition or absence of specific illustrative materials like graphs, maps, etc. Variations in abbreviations were ignored.

- Variations in Series Statement

  These include any difference in the statement of series such as order of terms in the statement. Variations in the abbreviation for "number" and "volume" were ignored.

4.2 Differences in Description

In the comparison of the 250 outside reports and LC replacements, 374 substantial variations were detected: 237 in the body of the entry (title through date of publication) and 137 in the collation (pagination through series statement). Thus, 63.4 percent of all differences occurred in the body and 36.6 percent occurred in the collation. The mean number of differences per record was 1.496. Within the body of the entry, the following differences were noted:

Variations in Title	13.1 percent
Variations in Author Statament	28.3 percent
Variations in Edition Statement	5.0 percent
Variations in Place or Publisher	35.0 percent
Variations in Date of Publication	18.6 percent
Total Variations in Body of Entries	100.0 percent

Within the collation area, variations were noted in the following categories:

Variations in Pages or Volumes	40.1 percent
Variations in Illustration Statement	46.7 percent
Variations in Series Statement	13.1 percent
Total Variation in Collation (rounded)	99.9 percent

Differences between the descriptive elements of the 250 records compared were extensive: 80.4 percent of all records had one or more variations. The descriptive elements with the largest number of variations were the place and publisher, the author statement, and the illustration statement. These constituted 57.2 percent of all differences noted. Variations in the place of publication alone were few, constituting only 4.5 percent of all the variations in the sample. The variations in the publisher statement alone constituted 13.4 percent of the differences between all descriptive elements, while combined variations on a single record of both place and publisher constituted 4.3 percent of all variations. Variations in the author statement resulted in 17.9 percent of all differences, while the illustration statement accounted for 17.1 percent of the differences in the sample. While these elements constitute a major portion of all variation, their significance in eliminating duplication from a union catalog is small. Variation in the author statement is compensated for by the comparison of entry headings associated with a record, which are more consistent in their form. On the other hand, variation in the illustration statement of the collation is not thought to be a useful discriminator for duplicates and is therefore not used in the checking for duplicates at the NUC. Recent studies show that automated removal of duplicates from data bases also ignores these data.3/ 4/ 5/ One automated method ignores the publisher entirely.6/

Differences in pagination constituted the fourth most common source of variation, accounting for 14.7 percent of all differences in the sample. The majority of these differences, however, were in the roman numbered preliminary paging. The importance of variation in paging is diminished for duplicate searching if only arabic numerals are used when in removing unwanted duplication in a file.7/

Variations in the series statement and edition statement resulted in the smallest number of differences: 8 percent. Differences in the series statement resulted in 4.8 percent of the variation, while differences in the edition statement constituted 3.2 percent. The occurrance of these fields in MARC records is low, as reported in a sample taken by Williams and MacLaury,8/ and edition comparisons were eliminated in searching for duplication in their Illinois Union Catalog study. These variations are therefore less important as discriminators of duplicates than search keys built on the title, which has a 100 percent incidence of occurrence.

The variations due to title resulted in 8.3 percent of all differences detected in the sample. The variations in the date of publication caused 11.8 percent of all differences. These two fields are of primary importance in the removal of duplicates from a file, as reported by the manual searchers in the NUC Control File and as indicated in recent literature reporting experiments with the automated

removal of duplicates.9/ These fields are also the most successful and useful points of access and means of discrimination between records for users of library catalogs, as reported in a synthesis of catalog use studies by Lancaster.10/ Most of the variations in the title field were due to the addition of subtitles, the completion of lengthy titles, and title abbreviations occurring in foreign cataloged works. Most of the variation in the date field occurred when minor changes were made in the date information (the addition of "c" for copyright date, or the supplying of later printing dates), although substantial differences resulted from the supplying of dates by LC that were unknown to the reporting library. A summary of all differences found in the sample is included in Table 7.

The average number of descriptive variations per record in this comparison was 1.514. The median number of changes per record was one. Although the number of differences reported is small, it must be noted that this report ignored differences in punctuation and in some abbreviations, as indicated at the beginning of this chapter. Of all the records, 36.1 percent had no variations within the body of the entry, while 39 percent had one difference, 21.6 percent had two differences, and only 3.3 percent had three differences in description within the body. No variations in the collation were found in 56.4 percent of the records, while 35 percent had one variation, 8.2 percent had two variations, and only one record had variations in all three areas (.4 percent).

The number of descriptive differences in the sample are given in terms of cumulative percent in Figure 6.

4.3 Comparison of Description and Heading Variations

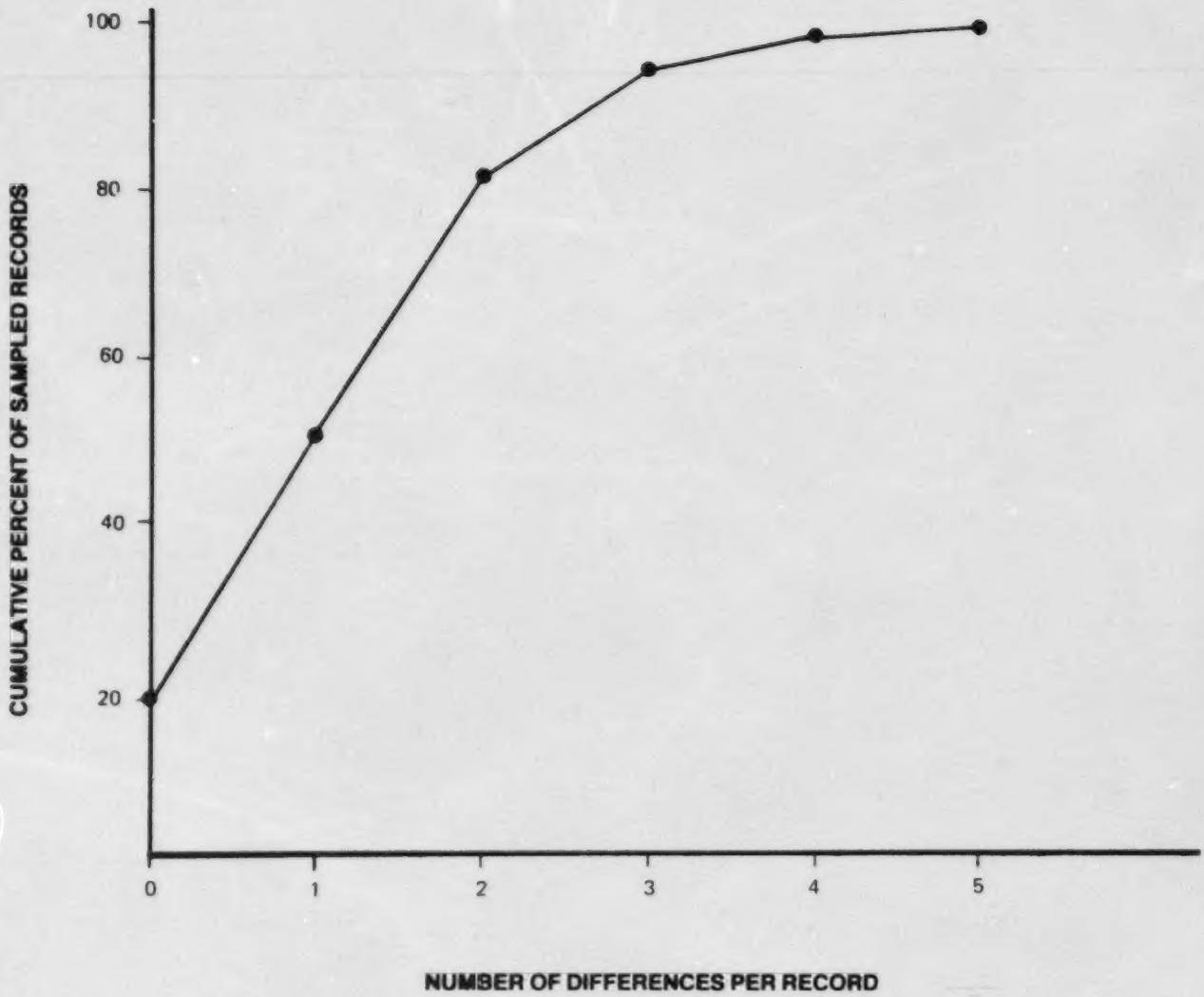
A number of different explanations may be given for cataloging variations between libraries. These range from the effect of local circumstances, including existing files, user needs, and staffing, to lack of the most recent information for cataloging, including name authorities and rule interpretations for descriptive cataloging. One question that could be investigated from the data available in this study was whether there existed a relationship between variations in descriptive cataloging and variations in headings, both entry and subject, between outside libraries reporting to the NUC and the cataloging records of LC. Is the variation in cataloging monodimensional, that is, do variations arise from a single source, such as the type of record being cataloged or the individual library cataloging the record? A Pearson correlation coefficient was calculated to determine whether variations in descriptive cataloging and variations in entry and subject heading work were related, and to what degree. The

TABLE 7

SUMMARY OF DIFFERENCES IN DESCRIPTIVE ELEMENTS
FROM COMPARISON OF LC AND NUC RECORDS
(N = 250)

AREA OF DIFFERENCE	NUMBER OF DIFFERENCES	PERCENT OF TOTAL DIFFERENCES
Title	31	8.3
Author Statement	67	17.9
Edition	12	3.2
Place and Publisher	83	22.2
Date	44	11.8
Pages and Volumes	55	14.7
Illustrations	64	17.1
Series Statement	18	4.6
TOTAL	374	100.0

FIGURE 6
BIBLIOGRAPHIC DESCRIPTION DIFFERENCES:
CUMULATIVE PERCENTAGE POLYGON
(N=250)



sample and statistics used were those reported in Chapters 3 and 4. The following standard equation for computing the correlation was employed:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$$

where:

r = Correlation coefficient
X = Number of differences in entry and subject headings on a sample record
Y = Number of differences in bibliographic description on a sample record
$\bar{X}$ = Mean number of differences in entry and subject headings per sample record
$\bar{Y}$ = Mean number of differences in bibliographic description per sample record
$x = X - \bar{X}$
$y = Y - \bar{Y}$

Thus since $\Sigma xy = 71.95$, $\Sigma x^2 = 1{,}058$ and $\Sigma y^2 = 302.5$, the following correlation coefficient was computed:

$$r = \frac{71.95}{565.73}$$

$$r = .1272$$

With a correlation coefficient this small, it was necessary to calculate a test to determine whether an r of .1272 could occur by chance in a sample of 250. The test of the hypothesis for no difference ($H_0 : r = 0$) was conducted at the .05 level (95 percent confidence) with 248 degrees of freedom. The critical value of the test was 1.97. The standard formula for the test is given as:

$$t = \frac{.1272 \cdot \sqrt{248}}{\sqrt{1-.0162}}$$

$$t = 2.0198$$

The computation of the test yielded a significant correlation of .1272. That is, an r of .1272 would be significant in 95 out of 100 samples of 250 from the population sampled. An r of this size, however, is generally considered to be a slight, almost negligable, relationship. The calculation of the coefficient of determination (r^2) gives us a better perspective of the strength of this relationship. In this

correlation, r^2 was equal to .016. The interpretation of this coefficient is that only 1.6 percent of the variance in descriptive cataloging is explained by the variance in entry and subject heading work. We can therefore dismiss the notion that some single factor, such as the outside reporting library's lack of adherence to standards or the difficulty of the individual record is the single great contributor to differences in cataloging. Instead, one must look to more complex explanations.

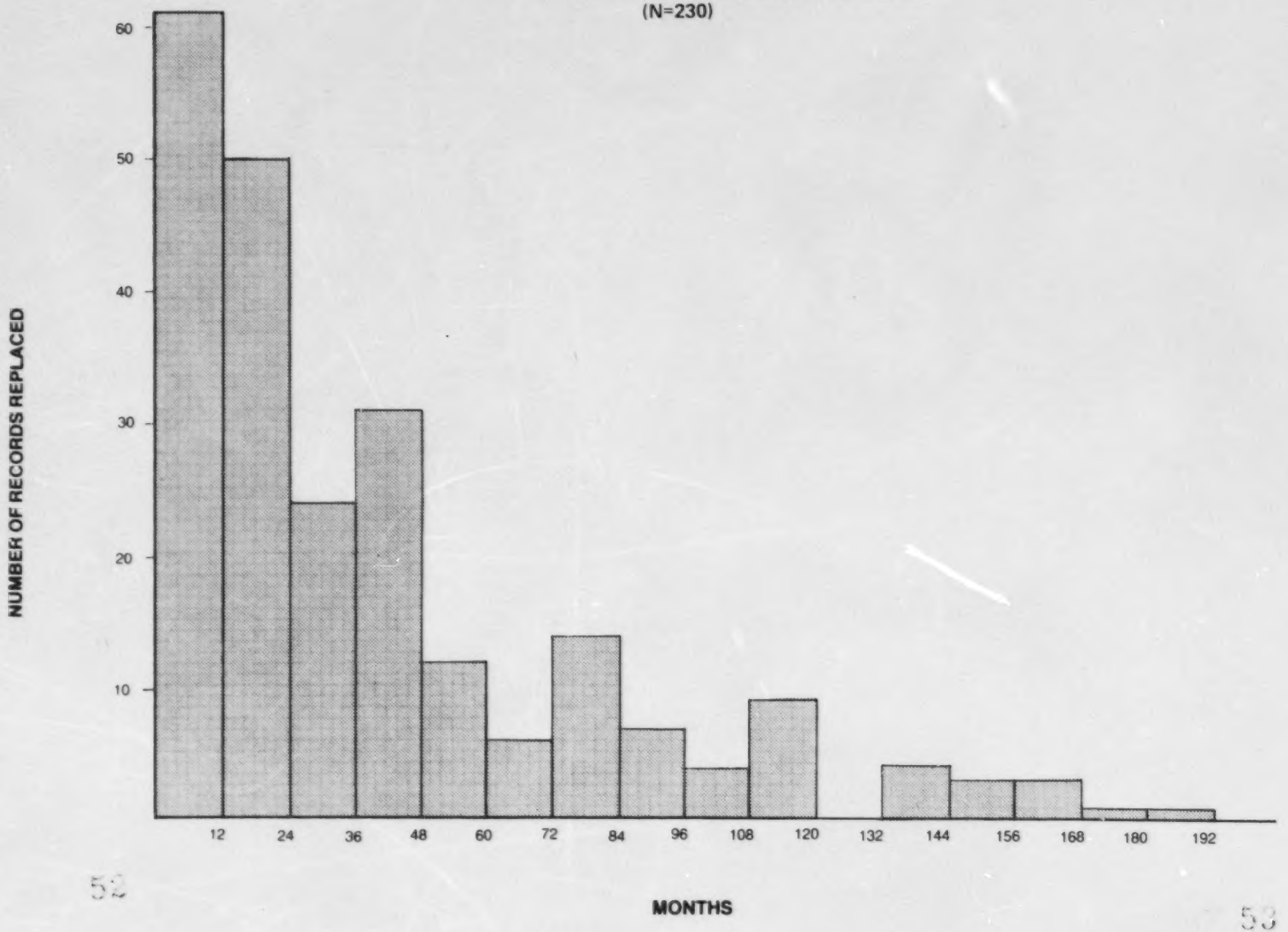
4.4 Time Taken to Replace NUC Records with LC Records

One factor to be considered in planning for a nationwide data base was the length of time it would take for a record transmitted by a member library to be replaced with a record of higher authority which would take precedence. To simulate this process, a random sample of 230 outside NUC reports which had been replaced by LC records was taken from the NUC Control File and the dates on which each record entered the Control File were compared. The interpretation of the time taken to replace NUC reports should be undertaken with a great deal of caution, because of the nature of the data collection. Although the dates on which each record entered the file (date of an NUC publication) were compared, the sample figure may be different from a "true" replacement rate, which the current operating condition of the NUC does not permit. A valid comparison would require both outside reports and NUC records to enter the Control File on receipt. Because of the backlog of outside reports to be searched, edited, and added to the Control File, outside reports may remain in process for as long as twenty-two months. Separate procedures are followed for editing and filing of LC records which place them in the Control File soon after they are received. It is possible, therefore, that while the comparison of dates in this sample may indicate a two-year period of replacement for a certain record, the actual elapsed time between the receipt of the outside report and the receipt of the LC replacement may be as long as four or five years. It may also be possible that an NUC report may arrive before an LC record and yet reach the Control File only after the LC record has already been entered. In this latter case, this sample would have revealed no indication of a replacement.

Figure 7 shows the distribution of replacements in relation to time using twelve-month class intervals. Intervals rather than raw data were used because the replacement dates corresponded to NUC publication dates, usually quarterly, and using precise months of replacement would lead to false precision. The distribution in Figure 7 is positively skewed, similarly to a Bradford distribution, which indicates a progression of replacements more logarithmic than arithmetic. The use of a mean value to characterize the "average" time taken to replace an outside report is misleading because of the presence of extreme scores. The median time of approximately thirty months is perhaps the best

FIGURE 7

NUC RECORDS REPLACED BY LC RECORDS IN
12-MONTH INTERVALS: FREQUENCY HISTOGRAM
(N=230)



"average" value and indicates that approximately one-half of those items to be replaced will be replaced in less than two and one-half years. It must be considered, however, that this "average" time for replacement may be in fact as long as five and one-half years.

Figure 8 shows the cumulative percentage of replacement in relation to time. It can readily be seen that over one quarter of the replacements occur within one year, and that 80 percent of all replacements are made within six and one-half years. This sample included only records that had been replaced. The determination of the number of items which are never replaced is thus not indicated and is beyond the scope of this report. Considering the rapid obsolescence of materials in many fields of knowledge and the need for quick access to materials in every discipline, the reports of outside libraries seem to provide a distinct and valuable supplement to the national collection available at LC.

REFERENCES

1. Paul S. Dunkin, Cataloging U.S.A. (Chicago: American Library Association, 1969), p. 49ff.

2. Arlene Taylor Dowell, Cataloging with Copy: A Decision Maker's Handbook (Little, Colo.: Libraries Unlimited, 1976), p. 38.

3. Thomas B. Hickey and David J. Rypka, "Automatic Detection of Duplicate Monographic Records," Journal of Library Automation 12 (June 1979): 125-42.

4. Keith D. MacLaury, "Automatic Merging of Monographic Data Bases--Use of Fixed-Length Keys Derived from Title Strings," Journal of Library Automation 12 (June 1979): 143-55.

5. Martha E. Williams and Keith D. MacLaury, A State-Wide Union Catalog Feasibility Study, Final Report on Project III-FY 1976 for the Illinois State Library (Urbana, Ill.: Information Retrieval Research Lab, University of Illinois, 1977).

6. Ibid.

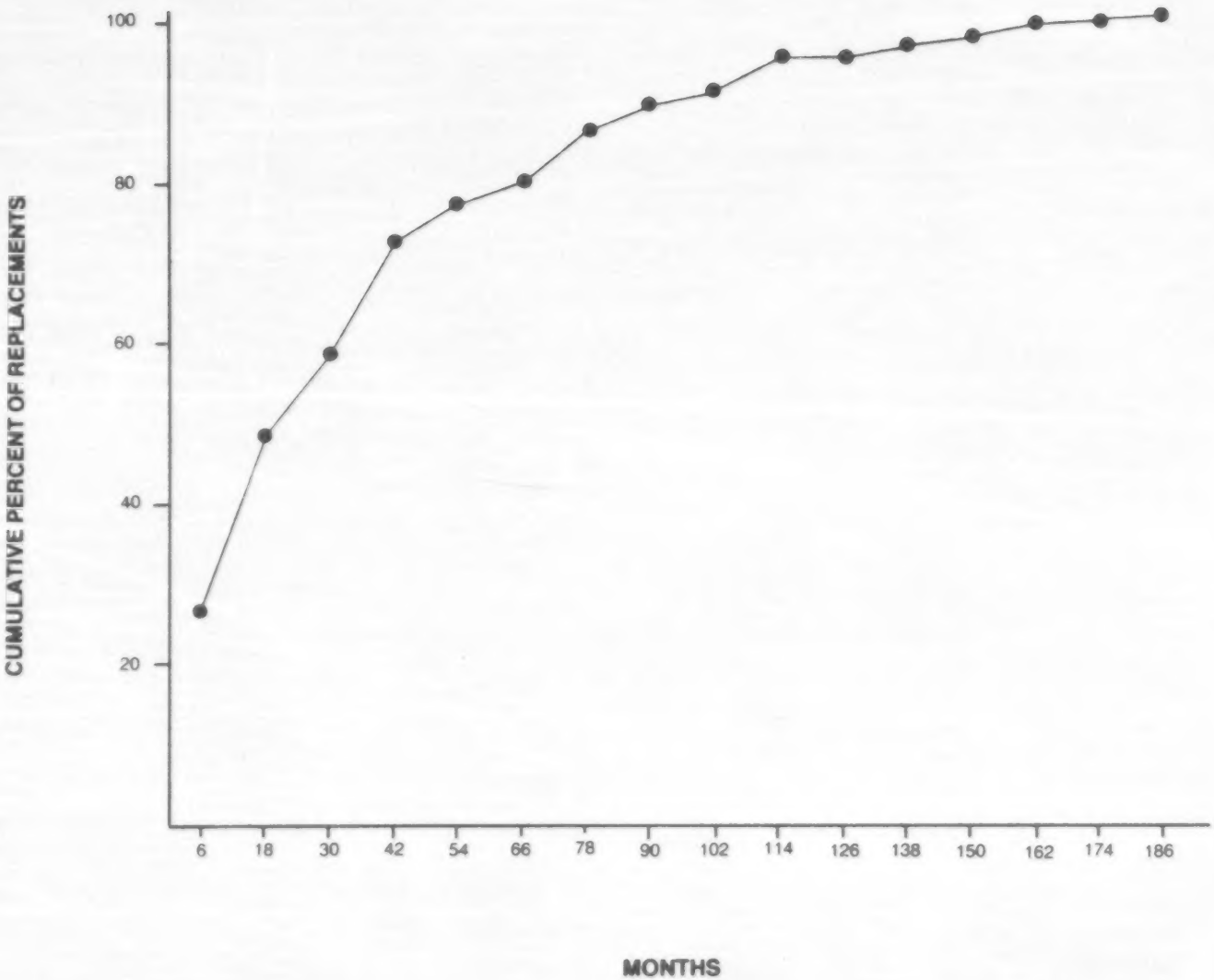
7. Ibid., p. 48.

8. Ibid., p. 17.

9. Hickey and Rypka, op.cit., pp. 128-30.

10. W. F. Lancaster, The Measurement and Evaluation of Library Services. (Washington: Information Retrieval Press, 1977), p. 70.

FIGURE 8
NUC RECORDS REPLACED BY LC RECORDS
IN-12 MONTH INTERVALS: CUMULATIVE
PERCENTAGE POLYGON
(N=230)



5. CONCLUSIONS

5.1 Conclusions for a Nationwide Data Base Derived from the Analysis of Processes in the Catalog Publication Division

The National Union Catalog has served libraries and the cause of scholarship in a unique and exemplary fashion since its inception in 1901. It provides valuable assistance in the areas of acquisition and collection development, cataloging, interlibrary loan, reference, and research to libraries throughout the world. The rich experience of the NUC can serve as both a guide and a stimulant to planning for a nationwide data base and network. The following conclusions are derived from a brief participation in certain elements of this experience and may be viewed as recommendations for future data base operations.

(1) A set of written guidelines must be created for the editorial sections of future centers of responsiblity and centers of special authorization to serve as a guide for editors and as a document to aid reporting libraries.

There is wide variation among NUC editors in their modification of outside reports, establishment of headings, and creation of cross references. The lack of current written guidelines that are consistently enforced permits editors to have the latitude for broad interpretation of general policies regarding the production of the NUC. This could lead to inconsistencies in a file which will inevitably diminish the utility of the product as users will be less able to anticipate potential access points in conducting their searches.

(2) The editorial staff of a proposed nationwide network should have the most current and open access to the policies and interpretation guidelines of the LC cataloging divisions in a usable form. This access should be broadened to eventually include all libraries that report their holdings to the data base.

Variations in cataloging records would be reduced if reporting libraries were able to incorporate LC rule interpretations in their cataloging. This incorporation would also reduce conflicts in local cataloging that are dependent upon LC and NUC reports for their cataloging data and potentially reduce cataloging costs on the local level as well as for the NUC Editorial and Control Sections. It can be anticipated that these variations and their resulting inefficiencies will exist in a nationwide data base unless coordinated action is taken.

Cumulative Book Index was less effective, with a verification rate of 32.8 percent. When this level of effectiveness is combined with the average time required to search titles, the NUC was rejected as an ineffective verification tool.

This lack of currentness can also be observed in the data derived from the study described in Section 2.1, which analyzed a sample of outside reports from the July-September 1976 quarterly edition of the NUC. No titles were found with 1976 imprints in the catalog, while 1975 imprints represented less than 1 percent of the records, and records of 1974 imprints were just being registered by 1976.

The diminished comprehensiveness of the NUC can also be observed in the comparison of growth statistics between the NUC and the OCLC data base. Figure 9 depicts the declining number of records published annually in the NUC. By contrast, the increase in records input and growth of the data base can be observed in the OCLC file, as Figure 10 and Figure 11 graphically demonstrate. The NUC Control File and the OCLC data base differ in the levels of effort expended in creating files with high bibliographic integrity and freedom from unwanted duplication. Operations oriented to machine-readable records and the exchange of information via telecommunication links would, however, expedite these processes for any future bibliographic data base and network.

5.2 Recommendations for Future Investigation

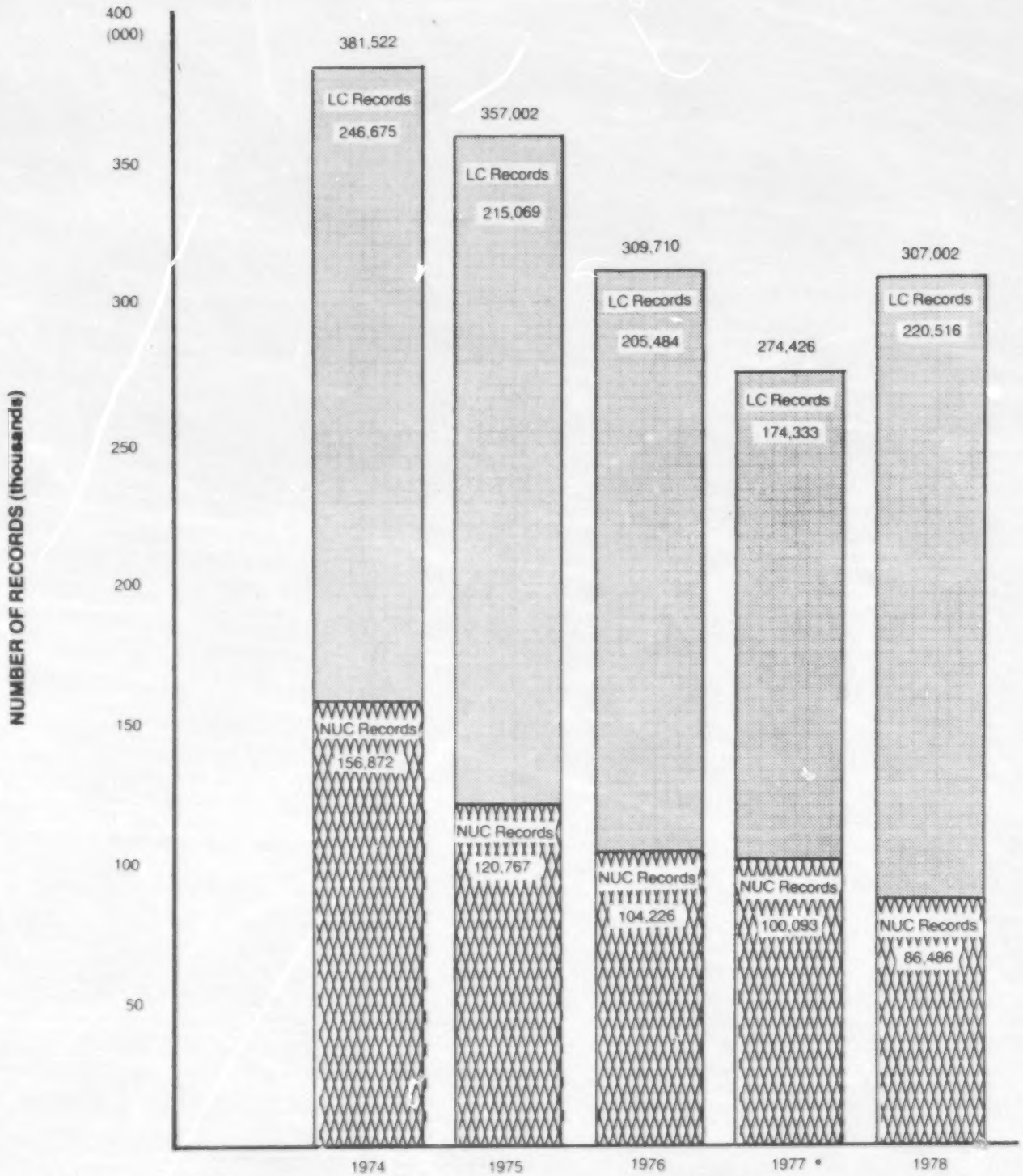
This study has established to some degree the relative rate of replacement of NUC reports by LC copy and the diversity of cataloging practice indicated by the extent of differences in the records. Each of these two problems still requires further investigation to ascertain a more precise rate of replacement independent of internal processing time and backlog for searching NUC reports and to determine the sources of differences in cataloging practice between libraries so that decisive steps can be taken to reduce inconsistent cataloging practices before an attempt is made to input records into a nationwide data base.

(1) Conduct a study in the form of a blind experiment to determine not only the differences between the records for documents cataloged originally by NUC contributors and LC but also the reasons for these differences.

The objective of this investigation would be to demonstrate causal links between specific cataloging conditions in NUC member libraries and variations between their records and LC cataloging records. This study, because of economic and time constraints, would have to sacrifice external validity (generalizability) for the internal validity which must be present if causal inferences are to be drawn. This study would therefore concentrate on an intensive investigation of

FIGURE 9

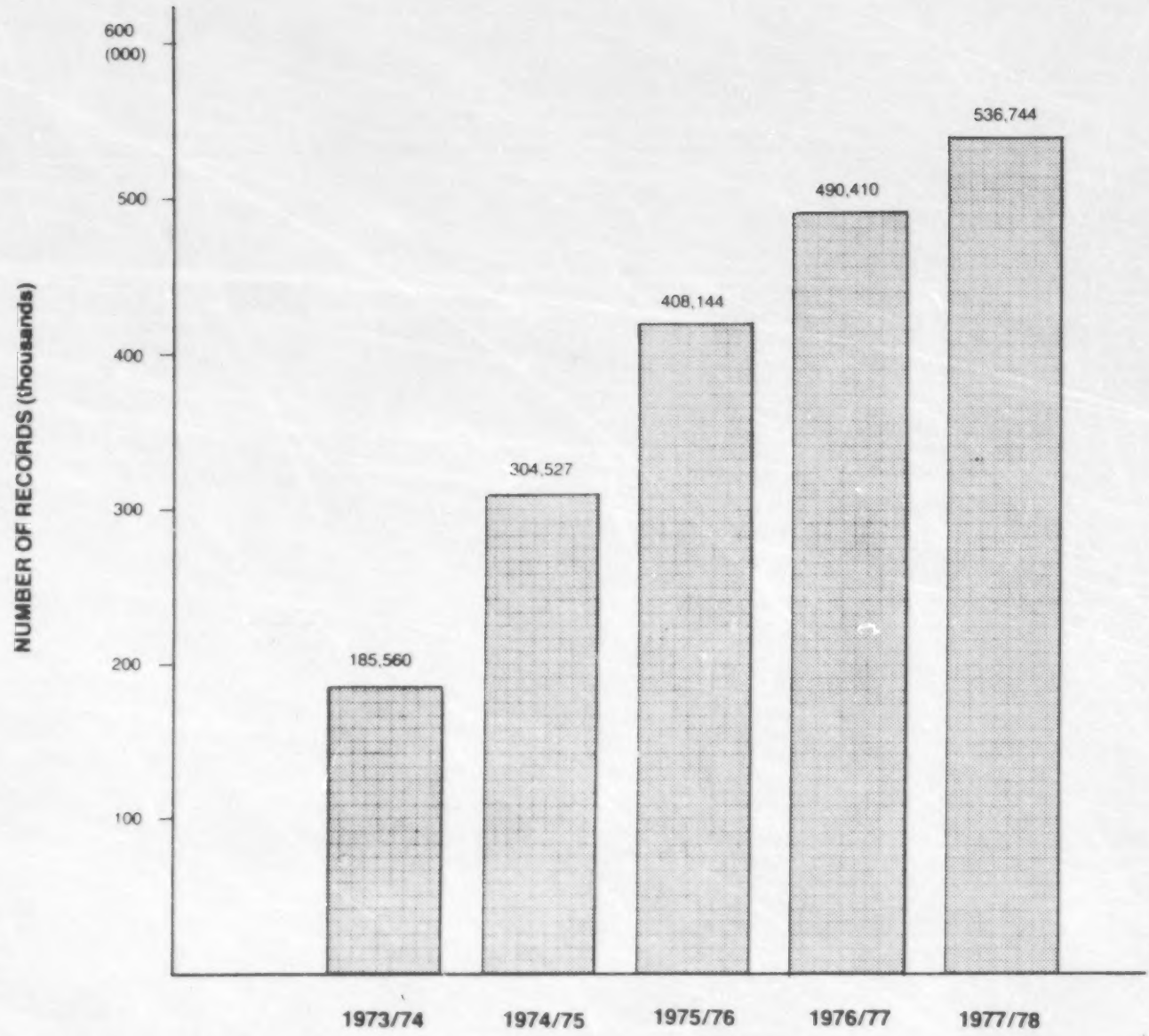
NUMBER OF RECORDS PUBLISHED ANNUALLY IN THE NUC



*1977 was the year for production of a 5 year cumulative NUC edition and the production for such years is frequently not typical.

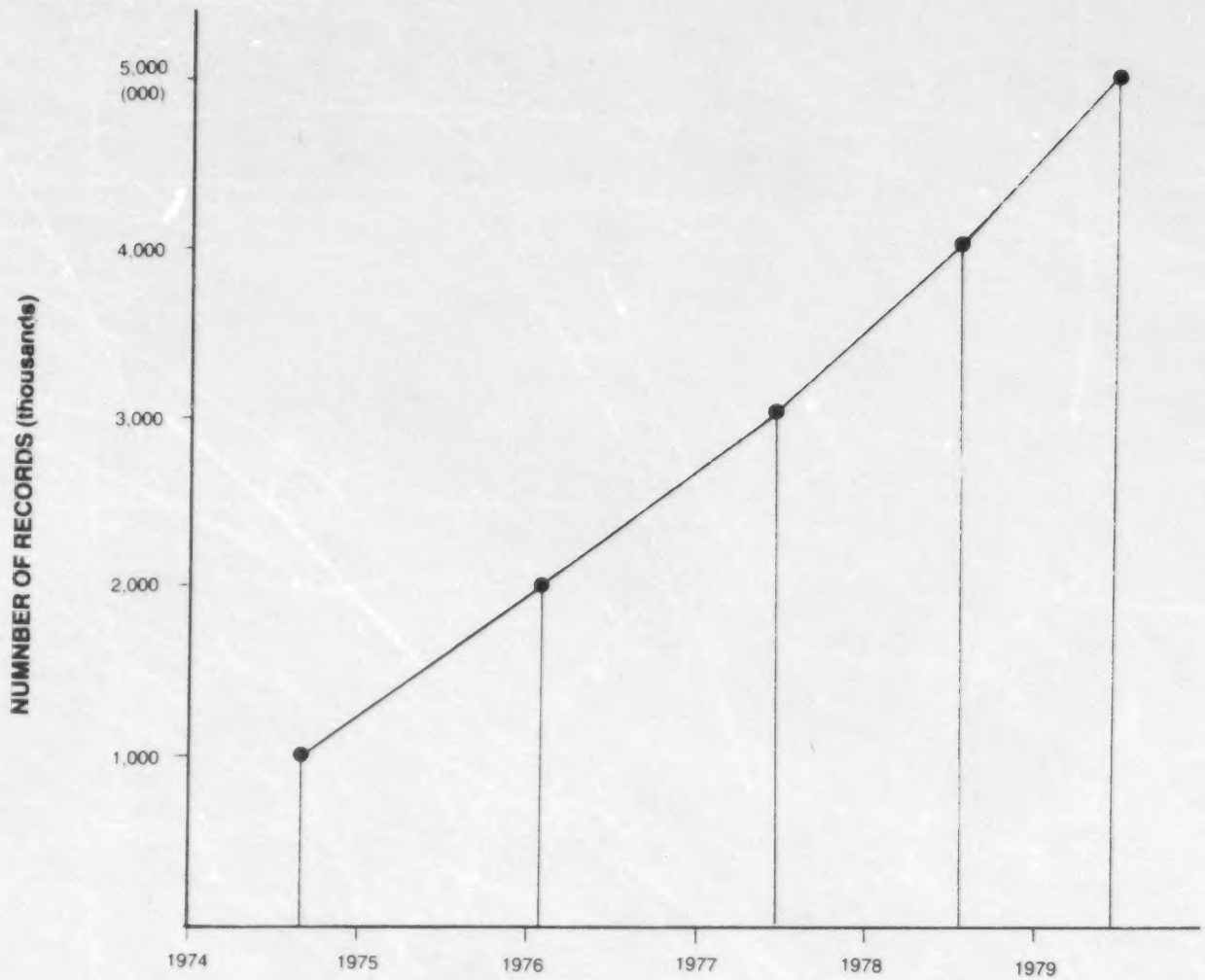
FIGURE 10

NUMBER OF RECORDS ADDED ANNUALLY TO THE OCLC SYSTEM



SOURCE OF DATA: OCLC, Inc., *Annual Report* 1973/74; 1974/75; 1975/76; 1976/77; 1977/78.

FIGURE 11
GROWTH OF THE OCLC DATA BASE



SOURCE OF DATA: OCLC, Inc., *OCLC Newsletter* No. 123 (June 13, 1979): 5

(3) Increase access to LC authorities.

The LC Official Catalog is the final arbiter for establishing authorities in the NUC Control File, and thus in the division's publications. This authority is currently imposed through the limited use of the Official Catalog by the NUC editorial staff and through the new LC records which establish new headings or serve to modify established ones. The physical distance of the Catalog Publication Division from Processing Services currently hampers the work of the Editorial Section. The inability of libraries reporting to the NUC to access the Official Catalog also leads to inconsistencies and differences in establishing headings in their reports. Planning for a nationwide data base must include access to all authority information in machine-readable form which can be shared through a telecommunication link.

(4) Any proposed nationwide data base must be designed for the receipt, processing, and dissemination of bibliographic information in a machine-readable mode to achieve a high level of cost-effectiveness in internal operations and to be of greatest utility in the delivery of services.

Of all records submitted to the NUC, 89.3 percent were from libraries which could have shared bibliographic information directly with the NUC in machine-readable form. Most of these records (85.6 percent) would be available through, and most of the libraries reporting (59 percent) would be members of, at least one of the three major bibliographic utilities: OCLC, WLN, and RLIN. The current manual processes of the Catalog Publication Division are cumbersome and slow. The processes of a proposed nationwide data base should take advantage of the current capabilities of reporting libraries that are no longer in a manual mode. Those libraries that cannot transmit bibliographic information in this form should be accommodated as a special system. The manual operations which are responsible for the approximately twenty-two month backlog of searching new outside reports at the NUC are currently diminishing the overall effectiveness of the NUC as a source of both location and bibliographic information. In a recent evaluation of the effectiveness of bibliographic services for verification purposes, the NUC demonstrates the effect of the internal manual operations responsible for its lack of currentness.

In the evaluation conducted by Reid,1/ the NUC was compared with other verification tools including OCLC, Inc., LC deposits, the American Book Publishing Record, and the Cumulative Book Index. Using a population of currently requested monographs in the English language with imprints ranging from 1974 to 1976, a sample of 534 titles was taken between May 17 and June 18, 1976. OCLC, LC deposits, and BPR all proved more effective than the NUC as a verification tool. Of all searched items, 92.7 percent were verified through the OCLC system, while only 44.6 percent were verified using the NUC. Only the

the records and internal conditions of a small number of libraries which were determined by judgment to be representative of libraries which would submit reports of catalog activity to a nationwide data base.

Ten libraries would be selected and approached for participation in this study. Each library would be requested to submit its original cataloging records and they would be searched to determine whether they have been cataloged by LC, are currently in process, or are in a low priority cataloging category which can be immediately cataloged for the purposes of this investigation. The individual libraries participating would be asked to retain their workslips and other precataloging records for all records submitted in this experiment.

Records would be compared for differences in all cataloging fields, and the results tabulated as in the descriptive portion of this study. The principle investigator would then schedule visits to each of the participating libraries, examine preliminary cataloging records and worksheets, and interview original catalogers to determine why specific decisions were made that varied from LC practice. The resources, policies, and guidelines for rule interpretation would also be examined at LC concerning these records to determine what factors were absent in the participating libraries which contributed to the variations in the cataloging records.

This procedure should supply the data necessary to determine the extent of reduced variation that would result from specific programs such as the rapid sharing of cataloging policy, guidelines, and interpretations which could assist libraries in establishing a higher degree of uniformity in cataloging practices. Factors which were beyond the control of a national program would also be identified.

(2) Conduct a study to investigate the rate of replacement and the time taken to replace individual library reports with LC records.

This investigation would be conducted using an on-line bibliographic utility such as OCLC which would not be subject to the processing delays and backlogs which now exist in the Catalog Publication Division. Currently input original cataloging would be sampled in the data base and monitored for replacement by LC records. Since it is a longitudinal study, this would take a minimum of one to two years to conduct. However, since examination of the NUC replacements within current operations does not permit a reliable measure of the records actually replaced, nor are specific time periods for replacement obtainable, this study would permit the only reliable measure of replacement rate and time.

REFERENCES

1. Marion Reid, "Effectiveness of the OCLC Data Base for Acquisition Verification," Journal of Academic Librarianship 2 (January 1977): 302,326.

ABBREVIATIONS

AACR	Anglo-American Cataloging Rules (1967)
LC	Library of Congress
LC MARC	MARC records created by the Library of Congress
MARC	Machine-readable cataloging
NUC	National Union Catalog (publication)
NUC	National Union Catalog (file)
RLIN	Research Libraries Information Network
WLN	Washington Library Network

OTHER NETWORK PLANNING PAPERS

No. 1 Butler, Brett. A Nationwide Location Data Base and Service. (Washington, Library of Congress, 1978). 66 p.

No. 2 Dataflow Systems. A Glossary for Library Networking. (Washington, Library of Congress, 1978). 34 p.

No. 3 Buchinski, Edwin J. Initial Considerations for a Nationwide Data Base. (Washington, Library of Congress, 1978). 56 p.

No. 4 Network Technical Architecture Group. Message Delivery System for the National Library and Information Service Network: General Requirements. (Washington, Library of Congress, 1978). 35 p.

No. 5 Long, Philip L. Study of Message Text Formats: Bibliographic Search Queries. (Washington, Library of Congress, 1979). 28 p.

Available by request from the Customer Services Section, Cataloging Distribution Service, Library of Congress, Navy Yard Annex, Bldg. 159, Washington, D.C. 20541.